Master Thesis - Msc Data Science and Marketing Analytics

What drives crowdfunding success?

A semantical analysis using descriptions of crowdfunding campaigns initiated in the US.

Author:
Dennis van Honschoten

Student number:
401732

Supervisor:
Dr. R. Karpienko

Second assessor:
Dr. M. van de Velden

Abstract

This thesis assesses the impact of possible campaign success drivers using a semantical approach. Mainly, reciprocity and altruism are discussed as drivers of donation intention. Crowdfunding campaigns initiated in the US are collected from the online crowdfunding platform *GoFundMe*. The results of this thesis suggest that utilizing altruistic cues and positive polarity in campaign descriptions lead to higher donation amounts. Also, electronic Word-of-Mouth is found to have a positive impact on donation amount. Finally, reciprocal cues are found to be significantly and positively influencing campaign success. This indicates that utilizing reciprocal cues in campaign descriptions can help campaign initiators to 'make the campaign happen'.

Date final version:
December 22, 2019

Table of contents

1. Introduction

Crowdfunding has risen in popularity after the financial crisis of 2008 as a way for entrepreneurs to get access to monetary funds (Lee et al., 2010). As of 2019, the total transaction value in the crowdfunding sector is expected to be around 6.8 billion dollars for over 8.6 million campaigns held (Statista, 2019). Belleframme et al. (2013) define crowdfunding as "raising external financing from a large audience (the "crowd"), in which each individual provides a small amount, instead of soliciting a small group of sophisticated investors".

The risen popularity of crowdfunding and technology such as the world wide web have allowed for crowdfunding platforms to emerge. In 2015, around 450 crowdfunding platforms were known to be active (Cordova et al., 2015). These platforms allow organizations and individuals in search of funds to connect with potential donators by formulating their story and have shown to mitigate some of the drawbacks of traditional funding, such as the effect of distance between the entity in search of funds and the provider (Agrawal et al., 2011). Although each platform tend to focus on a different sector (for example, Kickstarter's phrases its main mission as "to help bring creative projects to life"), most crowdfunding platforms allow both non-profit and for-profits campaigns to be listed. This thesis will focus on identifying those aspects that drive campaign success for both forms of crowdfunding. For both for-profit and non-profit crowdfunding, literature has divided the different methods that an initiator could apply in order to reach their target amount.

Within the context of for-profit organizations, crowdfunding is most often used as a way for start-ups to acquire seed capital. In a similar fashion, firms use the received capital for the funding of innovations or other investments (Schwienbacher and Larralde, 2010). Acquiring capital using crowdfunding has a significant benefit over traditional funding, because there is no financial intermediary involved. This lowers the transaction costs (Lam and Law, 2016). Other than using crowdfunding to acquire capital, entrepreneurs can establish a crowdfunding campaign to assess the demand of their product. This can in turn be used as a tool to acquire capital in other ways, such as venture capital funding. Additionally, crowdfunding could be used as a marketing mechanism in order to attract press attention (Mollick, 2014).

Literature on for-profit crowdfunding distinguishes two major forms of crowdfunding (Belleflamme et al., 2014). The first form is *pre-ordering*, in which individuals are able to pre-order the product in order for the company to have enough capital to launch the product. These pre-purchasers then receive the product once production has been finished, or are given the right to buy the product at a reduced price

(Griffin, 2012). This form of crowdfunding provides companies an opportunity to differentiate in price between the pre-purchasers and consumers that buy the product once it has been fully launched. This allows entrepreneurs to extract larger revenues. The second form of crowdfunding is *profit sharing* (or, *equity crowdfunding*). Investors are asked to provide money in exchange for equity securities of the company or some sort of share in future profits. Initially, this form of crowdfunding was banned in the United States since it involves offering securities without registering at the Securities and Exchange Commission (SEC) (Griffin, 2012). However, it was later legalized under the *Jumpstart Our Business Startups* (JOBS) act (2017). Other than these two forms of crowdfunding, companies may apply a *lending model*, in which contributors are asked to loan money which is repaid, mostly with interest, at the end of a scheduled period (Mollick, 2014). Finally, a company may opt to apply a *reward model* in order to attract customers. Using this model, funders are rewarded with 'thank-you' gifts associated with the company, like signed t-shirts or an acknowledgement in the product description. This is also called *patronage crowdfunding* (Burkett, 2011).

As for non-profit crowdfunding, a majority of the campaigns are set up in order to collect funds to finance a certain project. This could either be for the restoration of a building, a medical intervention for a person in need or to cover the costs for a sports team to participate in an important competition. As the name suggests, this differs from for-profit crowdfunding in the sense of that a participant in most cases receives no monetary incentive from donating to the campaign. However, non-profit campaigns and their success are mainly driven by altruism and psychological factors such as empathy.

The differences between NPO and for-profit donating behaviour seem to be driven by the motivation behind an individual donating. Venable et al. (2005) describe in their paper that for-profit donations are mostly driven by a monetary-based exchange, whereas NPO donation behaviour tends to be more socially driven. Accordingly, Lam and Law (2016) state that for-profit donating are mainly based on extrinsic motivation of the donator, whereas NPO donating is mainly based on intrinsic motivation. However, for-profit donating motives shares some similarities with NPO donating, such as the feeling of doing the right thing and identification with the organization (Harms, 2007; Shehu et al., 2016) and enhanced reputation (White and Peloza, 2009; Vesterlund, 2006).

This thesis tries to identify those aspects that makes campaigns for both forms of crowdfunding successful using a semantical approach. In the models used to predict campaign success, both numerical and semantical variables will be used. Therefore, the following main research question has been formulated:

The different methods being used in order to answer the research question will be explained in the remainder of this thesis. This thesis will apply different Natural Language Processing (NLP) techniques to create semantical input variables, such as Latent Dirichlet Allocation (LDA) and Global Vectors (GloVe). Logisitic- and linear regression will be used for the answering of hypotheses in this thesis.

This thesis will add to the existing literature for different reasons. First, predicting campaign success using semantical features has previously only been done for for-profits organizations such as start-ups (see for example Yuan et al., 2016). This thesis considers both for-profit and non-profit crowdfunding, by extracting data from the popular crowdfunding platform *GoFundMe*. Additionally, this study will use the novel word embedding method *GloVe* (Pennington et al., 2014) in order to build lexicons for concepts that explain donation behaviour for both for-profit and non-profit campaigns, as will be discussed in sections 2.1 and 2.2. These lexicons will be used to identify altruistic and reciprocal cues used in campaign descriptions.

This thesis aims to distinguish possible benefits of exerting charitable behaviour. For each of these benefits, its effect on campaign success will be assessed by transforming them into input variables. This will be discussed in sections 2.1 and 2.2. Using existing literature, drivers of successful campaigns for both non-profit and for-profit crowdfunding will be identified in section 2.3. This section forms the foundation of the control variable selection used in this thesis. Furthermore, brand identification and word-of-mouth (WOM) will be discussed as concepts which could potentially be of significance for predicting campaign success. To summarize, the first four parts of the literature review will consist of:

1) Intrinsic motivation and altruism (section 2.1);
2) Extrinsic motivation and reciprocity (section 2.2);
3) Drivers of campaign success (section 2.3);
4) Brand identification and word-of-mouth (section 2.4).

Section 2.5 will discuss topic modelling and its contribution for predicting campaign success. Finally, section 2.6 will discuss literature how the findings of section 2.1 and 2.2 can be used as input to predict campaign success for both for-profit and non-profit crowdfunding. This will involve creating concept-specific lexicons in order to identify altruistic- and reciprocal cues used in campaign descriptions. Furthermore, sentiment analysis will be discussed in this section because it could provide additional

(semantical) insights in predicting campaign success. The remainder of the thesis will consist of the methodology, data, results, conclusions and limitations.

2. Literature review

As discussed, this section of the thesis will first be split into four streams of research that explain donation behaviour. Sections 2.1 and 2.2 focus on altruism and reciprocity as being motivators for individuals to donate to a crowdfunding campaign. Furthermore, differences are expected for using reciprocal versus altruistic cues in non-profit crowdfunding. This will be discussed in more detail in section 2.2. Section 2.3 provides an overview of literature that focusses on identifying drivers of successful crowdfunding campaigns for both non-profit and for-profit crowdfunding. This section forms the foundation of the control variable selection used in this thesis. Additionally, section 2.4 discusses the effect of word-of-mouth for crowdfunding campaigns. Section 2.5 focuses on topic modelling and its relevance in identifying semantical insights from campaign descriptions. Finally, section 2.6 discusses the means in which polarity used in campaign descriptions could influence campaign success.

2.1 Explaining donation behaviour: Intrinsic motivation and altruism

Literature that focuses on the rewards that a donator receive from participating in for-profit and non-profit crowdfunding can be divided into two sectors, namely rewards in the form of *intrinsic motivation* and *extrinsic motivation* (Kleeman et al., 2008; Schwienbacher & Larralde, 2010). Section 2.1.1 will discuss intrinsic motivation and its connection to altruistic behaviour (which will be discussed in section 2.1.2).

2.1.1 Intrinsic motivation

Intrinsic rewards can be described as psychological gains, such as enjoyment (Deci and Ryan, 1985). Previous crowdfunding literature has identified intrinsic rewards that significantly predict donating intention. For example, Harms (2007) found self-expression and enjoyment as significant predictors of donation intention. Furthermore, Van Wingerden and Ryan (2011) found that for individuals that are driven by intrinsic benefits, control of use over an innovation, a sense of involvement and enjoyment are the primary drivers of donation. By using controlled experiments, Zvilichovsky et al. (2018) found that motivation to participate in crowdfunding is mostly product-centred and not people-centred. In other words, they found that the motivation is driven by a desire to make the product happen and not necessarily by a desire to help the entrepreneurs.

Within the context of prosocial behaviour, intrinsic motivation has been linked to altruistic behaviour. For example, Warneken and Tomasello (2009) found that individuals who received a monetary reward prior to helping an individual were less likely to engage in further helping than individuals who did not

receive a reward beforehand. They conclude that individuals have an intrinsic motivation to perform altruistic behaviour. The psychological perspective of altruism uses intrinsic rewards as a fundamental assumption in explaining altruistic behaviour. The next section will elaborate more on the different views of altruism and its role in explaining donation behaviour.

2.1.2 Altruism

Altruism has been linked to charitable giving in numerous papers. Altruistic behaviour can be described as behaviour that benefits another entity, while being damaging for the entity performing the behaviour and is characterized by these two entities not being closely related to each other (Trivers, 1971). Some authors argue that altruism can be explained from two different perspectives; one being the evolutionary perspective, and the other being a psychological perspective (Wilson, 1992; Rose-Ackerman, 1997). Furthermore, altruism can be explained from an economic point-of-view, by using an utility-maximizing approach (Andreoni, 1990; Sugden, 1982).

First, the evolutionary perspective of altruism is discussed in depth by Sober and Wilson (1999). This perspective concerns the effects of altruistic behaviour on evolutionary concepts such as reproduction and survival. They argue that human nature has been evolved to do what is for the benefit of the group. Sober and Wilson describe this as the *within-group adaptation*. From an evolutionary perspective, those individuals that help other individuals to survive and reproduce at their own expense will be eliminated based on natural selection. However, the authors argue that this differs on group-level. Within groups, individuals will adapt towards group-morality in order to survive when conflicts (such as war) arise.

The psychological perspective of altruism is based on motives of the actor (Wilson, 1992; Clavien and Klein, 2010). These motives can be driven by various processes. For example, De Waal (2008) describes empathy, the perception of the emotional state of another individual, as an important driver of altruistic behaviour because it triggers the individual's own emotional state. By analysing brain activity, Decety et al. (2015) conclude that the ability to empathize stimulates prosocial behaviour. Other motives of charitable giving include peer pressure (Sugden, 1982; DellaVigna et al., 2012), an ambition for income equality, or a belief that the public sector provides insufficient funds to a particular sector (Rose-Ackerman, 1996).

The economic literature on altruism shares similarities with the psychological approach: donors may gain utility from knowing that their donation made someone less fortunate better off (Becker, 1974).

Using insights from the paper of Becker (1974), Sugden (1982) introduced his version of the *public good theory of philanthropy*. According to this theory, an individual maximizes his utility not only by distributing his income between personal consumption and charitable contributions, but also by taking the actions of other donators in consideration. Andreoni (1989; 1990) referred to this as *pure altruism*, in which an individual's motivation to donate is solely based on the desire to donate to raise funds. This is distinguishable from *impure altruism*, in which an individual's motivation to donate is additionally driven by the joy he receives from donating. Impurity in this case refers to the egoistic aspect that plays an important role in explaining donation behaviour. Impure altruism has been introduced by Andreoni (1990) in his model of *warm glow giving*. According to his model, an impure altruist's utility function is given by:

$$U_i = u(x_i, G_{-i} + g_i, g_i)$$

For which x_i = i's consumption of private good, G_{-i} = contributions to the public good by everyone except i and g_i = i's contribution to the public good. Thus, an impure altruist not only receives utility from the total contribution to the public good, but also from the pleasure of the giving itself (or, *warm glow*). For a pure altruist, its utility function is as followed:

$$U_i = u(x_i, G_{-i} + g_i)$$

As shown, a pure altruist is only interested in the total contribution of the public good and not in the way in which the contribution is distributed between himself and others. As a result, G_{-i} and g_i are described as perfect substitutes. An increase in the amount donated by another individual reduces the amount donated by individual i by the same amount. This phenomenon is called *crowding out*. If the total contribution to the public good (G) is fixed, a pure altruist has no incentive to donate. This is because a pure altruist is only interested in the total contribution to the public good. For an impure altruist however, the 'warm glow' will result in the individual still donating towards the public good even in the case of complete crowding out.

In line with previously described literature about altruism and donation intention, this thesis will test if altruistic cues used in campaign description significantly impact campaign success and total amount funded. Accordingly, the following hypotheses have been formulated:

Hypothesis 1a: Altruistic cues used in campaign descriptions positively influences the probability of a campaign being successful.

Hypothesis 1b Altruistic cues used in campaign descriptions positively influences the amount of funds received.

2.2 Explaining donation behaviour: Extrinsic motivation and reciprocity

As discussed in the introduction of section 2.1, rewards from participating in crowdfunding can be divided into intrinsic and extrinsic rewards. Section 2.2 will focus on the latter and its connection to reciprocal behaviour.

2.2.1 Extrinsic motivation

Extrinsic motivation is driven by external incentives, such as rewards and recognition from others (Deci and Ryan, 1985). By examining each form of crowdfunding for for-profit firms (as discussed in the introduction of this thesis), we see that each form could include some sort of extrinsic reward for the donator. This can either be a share in the future profits of the firm, interest received on issued loans, a pre-ordered product or a 'thank-you' gift. Using questionnaire-based research, Harms (2007) found the perceived economic value (the overall benefit a donator received with respect to his contribution) and the availability of guaranteed tangible output as the strongest predictors of donation intention. Additionally, Van Wingerden and Ryan (2011) found two types of donators, those that either participate in crowdfunding to receive extrinsic rewards or intrinsic benefits. For those that participated based on extrinsic rewards, monetary rewards are the most important driver. Similar as for non-profit donations, White and Peloza (2009) conclude that reputation improvement also is a significant driver of for-profit donation intentions.

Extrinsic motivation is closely related to reciprocal behaviour, because an individual could perform a reciprocal act in order to receive some sort of reward in the future. The next section will discuss reciprocity and its applications within different streams of literature.

2.2.2 Reciprocity

Closely related to extrinsic motivation, reciprocity can be defined as a situation in a social environment in which two individuals or organizations perform a mutual exchange, which could either be rewarding kind actions (*positive reciprocity*) or punishing unkind actions (*negative reciprocity*) (Gouldner, 1960). Similar as for altruism, literature distinguishes different perspectives of reciprocity. These will be explained using an evolutionary, psychological and economic approach.

First, the evolutionary approach considers the development and the survivability of reciprocity. According to Buunk and Schaufeli (1999), three conditions must be present in order for reciprocity to survive. First, the environment must provide sufficient opportunities for favourable cost-benefit ratios of reciprocal actions. That means, favours towards others should cost little for the individual performing the reciprocal action and should provide a chance of a significant future benefit. For example, when meat-eating animals hunt for food, most often they kill prey and end up with too much food that they can consume at that time. Sharing food exerts little effort and when the animal is not able to hunt for several weeks, it could benefit significantly when the kind action is returned. Secondly, environment should provide sufficient opportunities for reciprocal actions to emerge. For example, long-term contact among individuals in an environment makes reciprocal favours more likely to occur. Finally, it is important that mechanisms exists that identify and punish cheaters. For example, in prison it is common practice to swap parts of a daily meal with other prisoners on separate days. In such an environment, prisoners who did not return the favour of giving away a part of their meal on a later day get punished verbally and physically by other prisoners. Another example is a tit-for-tat strategy in trust games, which will be discussed in the economic perspective of reciprocity.

Within the psychological perspective, researchers have focused on exploring the role of reciprocity not only as a personal belief, but also its function in social systems. Gouldner (1960) was one of the first to argue that reciprocity is part of almost every social system and only a few are exempted from it. The existence of reciprocity as a social norm is what he called the *norm of reciprocity*. As a social norm, reciprocity has been used in social studies to elucidate many different phenomena, such as organizational support (Settoon et al., 1996), international relationships (Keohane, 1996) and employee-employer relationships (Dabos and Rousseau, 2004).

From an economic perspective, reciprocity has been used to rationalize decisions made by consumers in experimental games, such as the trust game (Berg et al., 1995). In a trust game, two participants are both placed in a different room and given a fixed amount of money (let's say $10) as a show-up fee. The participant in room A must decide how much of their money (M_a) they send to their anonymous counterpart in room B. The amount participant A chooses to send to participant B gets tripled, which results in player B receiving $3M_a$. Participant B then must decide how much money to send back, which is denoted by $k_b(3M_a)$. This will result in a payoff for participant A equal to:

$$P_a(M_a, k_b) = \$10 - M_a + k_b(3M_a)$$

And for participant B:

$$P_b(M_a, k_b) = 3M_a - k_b(3M_a)$$

If a participant's utility is strictly increasing for wealth (W), its utility function is as followed:

$$U_i = W_i + P_i(M_a, k_b)$$

Maximizing U_i, participant B will keep all the money that participant A sends. Using backwards induction, participant A will choose to keep all the money. However, empirical evidence shows that participants often end up at a cooperative outcome in which $M_a > 0$ and $k_b(3M_a) > M_a$. In order for this to happen, participant A must deviate from the subgame perfect equilibrium and participant B must positively reward this behaviour instead of playing their dominant strategy. This is an indicator of reciprocal behaviour.

The previous paragraph gives an example of indirect reciprocity. Indirect reciprocity occurs when two individuals only meet once (Nowak and Sigmund, 2005). In that case, participant B can keep all money given by participant A without being punished for it in sequential rounds. This is different from direct reciprocity, in which two individuals repeatedly encounter. In contrast to indirect reciprocity, direct reciprocity allows for punishment of unkind actions and rewarding kind actions. Direct reciprocity is applied in repeated prisoner's dilemma (Nowak, 2006). Consider two players, who each can choose to "cooperate" or "defect". If they play a single game, their payoff matrix is given by:

$$\begin{bmatrix} R/R\ (3/3) & S/T(0/5) \\ T/S(5/0) & P/P(1/1) \end{bmatrix}$$

For which blue and red indicates the actions and payoffs for player one and two respectively. In this case, $T > R > P > S$. This means that if both players choose to cooperate, both receive a reward (R) equal to 3. If player one chooses to defect and player two chooses to cooperate, player one receives T and player two receives S. While the Nash equilibrium in this example is equal to both players defecting, the global optimal would be for both players to cooperate, because $2R > T + S$. Axelrod (1987) simulated a tournament of repeated prisoner's dilemma and concludes that *tit-for-tat* (TFT) is the most optimal strategy in the absence of noise. This strategy involves choosing to cooperate in the first round and mirroring the actions of the other player in subsequent rounds. However, in a situation where noise appears (for example, a player mistakenly chooses the wrong option), TFT has problems correcting the mistake (Imhof et al., 2007). If player one erroneously chooses to defect while player two cooperates, both players will end up alternating between cooperating and defecting. In this way,

the global optimum will not be reached. Literature has developed different strategies to combat the possible noise in repeated prisoner's dilemma, such as *win-stay, lose-shift* (Nowak and Sigmund, 1993).

Reciprocity has been applied in many empirical studies to identify behaviour patterns, including effort exerted by employees (Fehr and Schmidt, 2006; Gneezy and List, 2006) and crowdfunding. Zvilichovsky et al. (2015) found that initiators of campaigns that previously supported others receive more funds from the campaign initiators that they supported (direct reciprocity) as well as from other individuals (indirect reciprocity). Additionally, Mitra and Gilbert (2014) analysed words, bigram, trigrams and phrases to identify the language used in Kickstarter campaigns that predicts successful campaigns. For these terms, they attempt to identify reciprocity and other social concepts. They found terms that significantly predict campaign success which could be linked to reciprocity (for example, they argue that "also receive two" indicates reciprocal intentions).

Reciprocal behaviour has not only been studied based on donations by the campaign initiator to other campaigns. In a recent study, André et al. (2017) explore the relationship between reward-based crowdfunding and theory of reciprocal giving and find some similarities. Most importantly, uncertainty of returning the favour. When a campaign does not meet its target, or for some other reason the project fails, the backer of the project does not always receive the gift that was promised to him. Additionally, uncertainty will exist about when the backer will receive his gift. This is similar to how reciprocity works in any other social interaction.

In line with previously described literature about reciprocity and donation intention, this thesis tries to identify reciprocal intentions using the campaign description given by the initiator and use these as input variables for the regression models. Accordingly, the following hypotheses will be tested:

Hypothesis 2a: Reciprocal cues used in campaign descriptions positively influences the probability of a campaign being successful.
Hypothesis 2b: Reciprocal cues used in campaign descriptions positively influences the amount of funds received.

Furthermore, differences are expected in total amount funded for using reciprocal cues in non-profit crowdfunding. Reciprocal cues used in campaign descriptions (such as rewarding donators with a gift) provide donators with an additional incentive to donate. However, providing monetary rewards could possibly lead to a crowding-out of charitable behaviour. The effect of monetary rewards on the intention to perform charitable behaviour has been studied previously. Mellström and Johanneson

(2008) empirically researched the theory by Titmuss (1970), who theorized that monetary compensation for blood donators reduces the incentive for them to donate. The authors of the paper found a significant crowding-out effect for female blood donators. In the context of crowdfunding, Shehu et al. (2016) studied the moderating effect of monetary incentives for brand identity dimensions on donation intention. Their findings suggest that monetary incentives given by NPOs for which the perceived trustworthiness is high have a negative impact on donation intention. In line with these findings, this thesis will test if reciprocal cues used in non-profit crowdfunding diminish the amount of funds these campaigns receive. Accordingly, the following hypotheses is formulated:

Hypothesis 3: An interaction effect is expected, such that reciprocal cues used in campaign descriptions is less positive for NPOs than for non-NPOs in terms of total amount funded.

2.3 Drivers of campaign success

Section 2.3 contains a literature overview of papers that explore the drivers of campaign success for both NPO and for-profit campaigns. This will be the foundation of the selected control variables for the regression models. Drivers of NPO crowdfunding success will be discussed in section 2.3.1. In a similar way, drivers of for-profit campaign success will be discussed in section 2.3.2.

2.3.1 Drivers of campaign success: non-profit organizations

Ferraro et al. (2005) found that donators that respond to an initial request (for example, answering questions that activate their emotional responses) are more willing to donate, because this decreases their self-regulatory resources. Furthermore, individuals are found to be more willing to help when a disaster occurs (Piliavin and Charng, 1990), when a charitable request involves family members and acquaintances (Stewart-Williams, 2007) and when individuals are shown the historical frequency of donations (Frey and Meier, 2004; Martin and Randal, 2008). Additionally, Martin and Randal (2008) used four settings in which the initial donation varied from nothing to 50 dollars. They found that the average donation of each setting was compliant with the initial donation amount. Accordingly, the average donation amount was the highest in the setting with highest initial donation amount. Using data from 50,000 crowdfunding projects, Pitschner & Pitschner-Finn (2014) compared the performance of for-profit and non-profit campaigns. Their analysis shows that non-profit projects are significantly more likely to reach their donation goal and receive a higher average donation amount compared to for-profit campaigns.

2.3.2 Drivers of campaign success: for-profit organizations

Zvilichovsky et al. (2018) studied crowdfunding for start-ups and donator's motivation to back the project. They found that campaigns that appeal to an all-or-nothing approach (that is, a project is only realized when the donation goal is reached) is a significant motivator for donation participation. Furthermore, Chen et al. (2016) studied the *appeal mode* of a campaign and its effect on donation intentions. They found that campaigns that appeal towards guilt has a positive and significant effect on donation behaviour. This shows that campaign descriptions that include content that triggers a certain part of the emotional state of an individual are more likely to receive donations. Other than that, they found that functional products received more funds than hedonic products (products with emotional or symbolic aspects). Lastly, they found that the amount of rewards levels has a negative correlation with donation intention for projects that include rewards. That is, the fewer reward levels a campaign contains, the more likely an individual will donate towards the project. Using the signalling theory as foundation, Mollick (2014) explored the effects of campaign quality on donation behaviour. He found that aspects that improve project quality, such as including a promotional video, minimizing spelling errors and the amount of updates, significantly impact donation intention. Furthermore, the influence of geographical distribution between the donator and the entrepreneur is discussed. Although Agrawal et al. (2011) conclude that crowdfunding mitigates many of the negative effects of this distance compared to more traditional funding methods, Mollick (2014) shows that geographic distribution can have an impact on donation intention. That is, project success is based on the fit between the project category and the nature of individuals living in the area in which a company resides and the presence of other companies that operate in the same product category near the area. For example, individuals with an occupation in either arts, design or entertainment are more likely to donate to projects within the creative category in their area. Stanko and Henard (2017) explored the market performance of companies that raised funds for their project using the Kickstarter platform. They find that not the amount of funds raised during a crowdfunding campaign, but the number of people that donate to their project significantly impacts the market performance of crowdfunded products. Furthermore, Cordova et al. (2015) found that an increase in the project duration and the amount of dollars contributed per day increase the success rate of a project.

As section 2.3 elaborates, many different numerical variables have been used in other papers in determining campaign success for both for-profit and non-profit entities. The numeric variables used in this thesis based on these findings will be discussed in section 4.1.1. These variables will act as control variables for the regression models.

2.4 Explaining donation behaviour: Brand identification and word-of-mouth

2.4.1 Brand identification

In marketing literature, numerous papers have been formulated that describe *consumer-brand identification* (CBI). These literature argue that brands are able to embody and communicate important values of humans, which allows consumers to identify themselves with these brands (Bhattacharya and Sen, 2001; Kim et al., 2001). CBI has been linked to concepts such as brand trust (Sung and Kim, 2010), brand loyalty (Kim et al., 2001) and brand love (Batra et al., 2012; Albert and Merunka, 2013).

Brand identification is suggested to play a crucial role in the evaluation of and the intention to donate to NPOs (Faircloth, 2005; Venable et al., 2005; Sargeant, 1999) . The capability of the NPO to project its 'personality' thus plays an important factor in this process. Venable et al. (2005) investigated whether individuals ascribe personality traits to NPOs and if they use these traits to differentiate between them. They found that this indeed was the case and additionally proposed four relevant dimensions of brand identification for NPOs: integrity, ruggedness, nurturance and sophistication. Building on this theory, Shehu et al. (2016) broadened the four dimensions of brand personality and explored the relationship of these dimensions and donation behaviour towards NPOs through the moderating role of monetary incentives given by NPOs. Although their paper mainly focuses on this moderating relation, they also found a direct positive relation between perceived honesty, reliability and nurturance (such as the NPO being loving and caring) by the donator and his willingness to donate.

2.4.2 Word-of-mouth

Within marketing literature, many conceptual frameworks have been established that comprise the antecedents and outcomes of brand identification. In this thesis, the focus will lie on those frameworks that describe the relation between brand identification and word-of-mouth (WOM). Specifically, brand love has been empirically shown to have a positive influence on WOM (Albert and Merunka, 2013; Batra et al., 2012; Kim et al., 2001). WOM has many different definitions, but most researchers agree that WOM consists of at least two aspects. First, WOM consists of communication about products or services between two individuals or groups which are independent from the company in question. Secondly, communication must be done through a medium independent of the company (Silverman, 2001). Traditionally, WOM has most often been measured using either inference or surveys (Godes and Mayzlin, 2004). Using surveys, researchers can directly identify WOM. For example, Bowman and Narayandas (2001) measure WOM by directly asking participants whether they have told others about the experience with a brand and, if they did, how many people they told. Inference techniques consist of analysing patterns of WOM without directly asking participants. Inference techniques have gained

popularity since the rising of online review platforms and social media. Accordingly, many researchers refer to this kind of WOM as electronic word-of-mouth (e-WOM).

Several researchers have used online reviews and ratings as a proxy of WOM (Dellarocas et al., 2004; Dellarocas and Narayan, 2006; Reza Jalilvand and Samiei, 2012). Crowdfunding platforms allow and encourage (potential) donators to share campaigns through social media. In this way, campaign creators can develop word-of-mouth opportunities which facilitates awareness (Moqri and Bandyopadhyay, 2016). Electronic word-of-mouth has in turn be shown to have a positive influence on the propensity to pay a higher price for a brand (Albert and Merunka, 2013) and purchase intention (Reza Jalilvand and Samiei, 2012; Fan and Miao, 2012). Based on these findings, it is expected that crowdfunding campaigns with high WOM are likely to receive more funds. This thesis tries to evaluate the impact of WOM on the total funded amount by using the number of Facebook shares of the campaign as an inference of electronic word-of-mouth. Accordingly, the following hypothesis has been formulated:

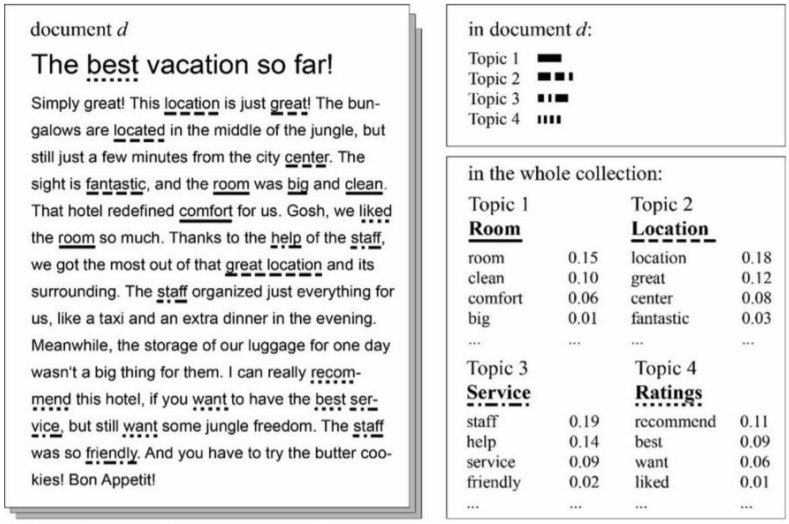
Hypothesis 4: An increase in Electronic word-of-mouth positively influences the amount of funds received.

To answer hypothesis 4, the number of Facebook shares will be used as a proxy for e-WOM and will be used as a input variable for multiple linear regression. Multiple linear regression will be discussed in section 3.3.

2.5 Topic modelling

Digitalization has had a major impact on our economy, with one of the outcomes being the availability of unstructured *big data*, such as large amounts of textual data (Kahn et al., 2010). An efficient way of analysing and understanding large amounts of documents is by expressing these documents in terms of underlying topics, otherwise known as *topic modelling*. Topic modelling is comparable with other dimensionality reduction methods (such as *principal component analysis* (PCA)) in a way that it tries to represent data using latent variables. This improves efficiency (because it takes less space to represent the data) and reduces noise (Rehurek and Sojka, 2010). Each of the topics can be interpreted by translating the word occurrence within each topic to a semantically correct concept. For example, figure 2.1 shows an example for a hotel review. Topic 1 consists of words such as “room”, “clean” and “comfort”. Thus, this topic could be described as aspects that describe a room at this specific hotel.

Figure 2.1: Example of topic modelling (Source: Reisenbichler and Reutterer (2019)).



Topic modelling and other semantical analyses are categorized as *natural language processing* (NLP) techniques. These techniques aim to capture human natural language by using artificial intelligence. This involves capturing meaning (or, *semantics*) from text. As trying to capture meaning is the core concept of NLP, one of the biggest issues is to determine a set of rules that tries to accomplish this. Nadkarni et al. (2011) explain the numerous challenges that NLP brings in capturing semantics, such as identifying relationships between words within a sentence, capturing part-of-speech (such as sarcasm) and emotions. In order for machines to understand language, one can assign specific rules. However, establishing an excessive number of rules could lead to rules hindering each other, which diminishes the original goal of NLP. In order to establish a coherent set of rules, *syntax techniques* have been developed. This refers to a set of rules that transform a sequence of words so that they make grammatically sense (Manning et al., 1999). Section 2.5.1 will discuss the techniques used in this thesis briefly. Applying these techniques allows campaign descriptions to be ready for the topic modelling technique used in this thesis (section 3.1) and the lexicon-based NLP techniques (section 2.6) in order to identify reciprocal- and altruistic cues and polarity used in campaign descriptions.

2.5.1 Syntax techniques

First, stemming is the process of transforming words by removing affixes, leaving the stem of a word (Manning et al., 2010). Closely related to stemming is lemmatization, which aims to capture the normalized form of the word (Plisson et al., 2004). In contrast to stemming, lemmatization tries to take the grammatical form of the word into consideration. For example, “meets” and “meeting” both get stemmed to “meet”, but as the former is a verb and the latter a noun, “meeting” should thus stay “meeting”. Finally, *part-of-speech* (POS) tagging aims to assign the correct POS to each individual word in the corpus (for example, “verb”, “adjective” or “noun”). Different algorithms exist that assign the

correct POS to each word. For example, Brants (2000) uses a *Hidden Markov Model* (HMM) to assign the most probable POS tag sequence for a given word in that sequence, dependent on the context probabilities and lexical probabilities over all POS tags. Additionally, Màrquez and Rodríguez (1998) transform POS tagging into a classification problem by using decision trees. In their analysis, they predicted the tag of a word using the tags of words within a window of 5 around the focal word.

After campaign descriptions have been cleaned for analysis using these syntax techniques, topic modelling can be applied. *Latent Dirichlect Allocation* (LDA) will be used in this thesis as a technique to identify topics given a corpus. In predicting campaign success, LDA can give additional semantical insights other than those obtained by creating a concept-specific lexicon for reciprocity and altruism. We can look at those topics that have significant impact in predicting campaign success and total amount funded and interpret these topics by looking at important terms within that topic. The technical background of LDA will be discussed in section 3.1.

2.6 Lexicon-based NLP techniques

In order to identify reciprocal and altruistic cues within campaign descriptions, a concept-specific lexicon will be built. This will be discussed in section 2.6.2. First, sentiment analysis is discussed as it also utilizes lexicons in order to assign polarity to documents.

2.6.1 Sentiment analysis

Within literature that focuses on utilizing NLP tools in order to conduct research, lexicon building has mostly been applied to capture sentiment within text. Sentiment analysis refers to classifying documents based on their polarity, which is either positive or negative (Pang and Lee, 2008). Sentiment analysis has most often been used to capture sentiment in product reviews, which in turn can be used to build consumer relationships (Homburg et al., 2015). Sentiment analysis has also been applied to analyse sentiment towards other entities than products, such as politicians and movies (for example, Mohammad and Yang (2011)). In order to capture sentiment, a specific lexicon has to be created that assigns polarity to each word. One of the most used sentiment lexicon nowadays is the *NRC emotion lexicon*. This lexicon assigns both polarity (either positive or negative) and one of eight emotions (anger, anticipation, disgust, fear, joy, sadness, surprise and trust) to unigrams. Assigning polarity and emotions is very time-consuming when done manually. As a solution, the developers of this lexicon (Mohammad and Turney, 2014) used crowdsourcing platform *Mechanical Turk* to divide the work into smaller parts and assigned them to people over the internet. As of today, their lexicon contains around 10,000 unigrams for which each a polarity score and emotion is assigned to.

Sentiment analysis can provide additional insights in predicting campaign success because campaign descriptions can contain different types of polarity. One example of polarity from the dataset used in this thesis is about crowdfunding to fund a documentary about Michael Jackson. This campaign focuses on all the negativity that was spoken about in the media about Michael Jackson and aims to raise funds in order to restore the image of Michael Jackson by creating a documentary about his life. This campaign uses primarily sentences with negative polarity, such as:

"..Once again, we have to defend Michael Jackson's name and legacy from vicious and calculated lies."
"..For the last 15 years of his life he was subjected to false, malicious allegations of child abuse."

Campaigns initiators can also opt to use positive polarity in describing the importance of their crowdfunding. Section 4.4 will discuss the way in which polarity will be used as an input variables for the dependent variables. Wang et al. (2017) investigated sentiment used in campaign descriptions for Kickstarter campaigns and conclude that positive sentiment is more likely to attract donators. Similarly, this thesis will test the following hypotheses:

Hypothesis 5a: Positive sentiment used in campaign descriptions positively influences the probability of a campaign being successful.
Hypothesis 5b: Positive sentiment used in campaign descriptions positively influences the amount of funds received.

2.6.2 Concept-specific lexicons

The traditional approach of manually building sentiment lexicons can be very time consuming. For this reason, researchers have applied word embedding methods to build concept-specific lexicons. This could be considered a semi-automatic approach, which reduces the amount of time necessary to build lexicons (Huang et al., 2014). Word embedding will be discussed in detail in section 3.2. Literature applying this method has been scarce, because lexicon creation using word embeddings is novel. One example of such research has been conducted by De Choudhury et al. (2013). Using a set of 900,000 questions and corresponding answers pairs of an online mental health forum, the researchers built a lexicon to identify depressive language. For each token in the document, the researchers calculated the association with the stemmed token "depress" using *pointwise mutual information* (PMI) and *log likelihood ratio* (LLR). Their lexicon finally consisted of the top-1,000 words with the highest *term-frequency, inverse document frequency* (TF-IDF) value associated with "depress". Applying word embedding methods for lexicon building has shown promising results. For example, Leroy et al. (2017)

compared the classification accuracy of manually created lexicons with those of automatically created lexicons for children diagnosed on the Autism Spectrum Disorders. These automatically created lexicons were built using *skip-gram* and for both lexicons accuracy was predicted using a neural network and a tuned classification tree. For both the neural network and the classification tree, the automatically created lexicons outperformed the manually created lexicons in terms of prediction accuracy (79.98% versus 76.92% using neural network and 86.81% versus 84.60% using decision tree, respectively).

In order to identify reciprocal/altruistic intentions captured in campaign descriptions, this thesis will build a lexicon for both concepts using automated word embedding algorithms. To build this lexicon, online articles that define and explain these concepts will be used. Another approach that could be considered is to consult online dictionaries, such as *Merriam Webster*, in order to extract synonyms of these concepts and use those to build a lexicon. However, such as approach is not optimal because it does not allow for distance to be a measurement of similarity between "reciproc" and "altruis" and tokens in their surroundings. Using a random sample of 100 campaigns, campaign descriptions will be manually scanned in order to classify them as either having reciprocal or altruistic intentions. This will be compared to the classification done by the concept-specific lexicon. One example of reciprocal intentions from a campaign within the dataset used in this thesis, which is about an individual aspiring to open a bakery store, is for example:

"*…Every donation will receive a personal thank you video."*
"…Donations of $10 or more, you get your own personal lifetime 15% discount code."
"…Donations of $100 or more, you get your own personal lifetime 25% discount code AND your name on the Broadway Baker wall of fame."

The goal of building the lexicons is to compute the similarity between reciprocity and corresponding tokens. Using the above example, "you get" would implicate a reciprocal action as a result of the initial donation. Our lexicons should be built in order to capture these reciprocal intentions. Section 3.2 will go in more detail about how the lexicons will be built using GloVe as a word embedding technique.

3. Methodology

Chapter 3 will focus on explaining the methodological background of the machine learning techniques applied in this thesis and the regression models. Section 3.1 will explain the means in which a set of topics are obtained from the corpora (which consists of the campaign descriptions). Section 3.2 will discuss word embedding and GloVe in order to build lexicons. Section 3.3 will focus on multiple linear regression and section 3.4 on logistic regression.

3.1 Latent Dirichlet Allocation

Latent Dirichlect Allocation, first introduced by Blei et al. (2003), is a unsupervised learning method which aims to capture topics using textual data as input. These topics represent latent dimensions of data and are characterized by a group of words that are available in documents. Furthermore, each document consists partially of all the hidden topics (Reisenbichler and Reutterer, 2019). This is different from classical clustering methods, where objects belong to a single cluster. For this reason, LDA is often referred to as a *soft clustering method* (Reisenbichler and Reutterer, 2019). Furthermore, LDA is regarded a *generative* model, in which it is assumed that documents are created by a (hidden) per-document topic distribution and per-topic word distribution (Rehurek and Sojka, 2010). Some of the benefits of this kind of topic modelling over others is that LDA can handle large amount of (textual) data and the ease at which the topics can be interpreted (Tirunillai and Tellis, 2014). Applications of LDA are used in different marketing- and other business related settings, such as customer satisfaction analysis (Guo et al., 2017), trend analysis (Moro et al., 2015), retail purchase prediction (Jacobs et al., 2016) and building recommendation systems (Christidis and Mentzas, 2013; Kim and Shim, 2014).

LDA has several assumptions. First, documents should contain a *bag-of-words*. This means that words appear independent given the documents, in such a way that the order of occurrence of the words in the documents does not matter (Wie and Croft, 2006). Furthermore, this method requires a vocabulary and a measure of word occurrence. Other than semantical applications, this technique also has been used for several computer visualization applications, such as image recognition (Wu et al., 2010) and localization (Filliat, 2007). Finally, the order of occurrence of documents in the data should not matter and the hidden topics should be set prior to the analysis (Blei et al., 2003).

Using LDA, we are trying to set the optimal number of K topics in corpus D, which is a sum of the amount of documents in the data: $D \in \{d_1, d_2, \dots, d_m\}$. Each document consists of $\{N_1, N_2, \dots, N_m\}$ amount of words, for which $w = (w_1, w_2, \dots, w_n)$ is the appearance of the words in the document, where w_n is the *n*th word appearing in the document. LDA models each of the documents in corpus D

as a combination of latent topics K, for which each topic describes a multinomial distribution over the entire vocabulary W in the corpus (Porteuos et al., 2008). LDA aims to represent the documents in terms of topics $\beta_{1:k}$, for which β_k is equal to the multinomial distribution of terms in topic K and $\beta_{1:k}$ will sum to 1 for all words belonging to topic K. Additionally, θ_d is the multinomial topic distribution per document and will sum to 1 for all topics belonging to a document. Prior to β_k and θ_d, LDA uses a Dirichlet to determine document-topic and topic-word distribution (α and η, respectively). To summarize, the generative approach of LDA is as followed (Blei et al., 2003):

1) Draw a word distribution for each topic $K \in \{1,2,\dots,K\}$ equal to $\phi_k \sim Dirichlet(\beta_k)$.
2) Draw a topic distribution $\theta_d \sim Dirichlet(\alpha)$ for each document d.
3) For each word N_m in the document:
   3a) Draw a topic assignment $z_{d,n} \sim multinomial(\theta_d)$ for which $z_{d,n} \in \{1,2,\dots,K\}$. Thus, $z_{d,n}$ represents the topic assignment per word for each document .
   3b) Draw a word from the selected topic: $w_{d,v} \sim multinomial(\phi_k = z_{d,n})$ for which $w_{d,v} \in \{1,2,\dots,W\}$ and $w_{d,v}$ is dependent on the topic distribution z_n in each document.

In determining the topic assignment per document, which is the hidden variable $z_{d,n}$, two posterior inference approaches are commonly applied (Nikolenko et al., 2017). The first approach is based on Markov chain Monte Carlo sampling. We want to assess the topic changes c given corpus D according to Bayes rule:

1) $P_{(z,c|w)} = \frac{P_{(w|z)}P_{(z|c)}P_{(c)}}{\sum_{z,c} P_{(w|z)}P_{(z|c)}P_{(c)}}$

Within this equation, $P_{(w|z)}$, $P_{(z|c)}$ and $P_{(c)}$ can each be solved using integrating (see Purver et al., 2006). However, the denominator has to be solved using Gibbs sampling. Gibbs sampling is an inference approach which allows obtaining topic assignments through an iterative process. The process is as followed (Steyvers and Griffiths, 2007):

1) Start with a random distribution of topics K for all words w in all documents in corpus D. This allows for an initial per-document topic distribution and per-topic word distribution.
2) Improve distribution by:
   a. Going through each word w in document d

b. Reassign word w_1 a new topic t, for which t is chosen based on the fraction of words N_m that are currently assigned to topic t in document d, multiplied by fraction of word w_1 assigned to topic t for all documents in corpus D.
c. Repeat for all words in all documents in corpus D.

So, through each iteration the previous topic assignment for word w is removed and updated based on the current per-document topic distribution and the current per-topic word distribution. This process continues until the topic distribution converges towards the true distribution (Steyvers and Griffiths, 2007).

The second approach involves variational approximations and was introduced by Blei et al. (2003). Using this method, the posterior distribution of the hidden variables are approximated by making use of Jensen's inequality to obtain an adjustable lower bound on the log likelihood (Jordan et al., 1999). A group of lower bounds, indexed by a group of variational parameters are considered at first. Then, the variational parameters are chosen based on a optimization procedure that minimizes the distance of the lower bound. This optimization procedure is described in more detail in Blei et al. (2003), as it is beyond the scope of this thesis. Porteous et al. (2008) and Hoffman et al. (2010) compared both variational approximation and Gibbs sampling and concluded that both have their advantages. Although Gibbs sampling is more accurate because it approximates the topic distribution more correctly, variational approximations are faster computed. This thesis uses Gibbs sampling to approximate the per-document topic assignment.

In setting the optimal amount of clusters, *perplexity* is often used as a measurement in semantic modelling (Jacobi et al., 2016; Yuan et al., 2016; Asuncion et al., 2009; Hoffman et al. 2010). Perplexity is mathematically equal to the inverse of the geometric mean per-word likelihood and is comparable to a goodness-of-fit measure for statistical models (Jacobi et al., 2016). For a range of values for K, we first train LDA on a part of the data and then evaluate the model using a set of test data. K will be set based on the model with the lowest value of perplexity. Section 4.3 will discuss the extraction of topics from the corpus and the way in which these topics will be used as control variables for the classification.

3.2 Word embedding for lexicon building

Theory about deriving meaning of a word within a context in NLP literature is divided into two streams: by using a symbolic approach and by using a distributional approach (Clark and Pulman, 2007). The

symbolic approach is most often used in deriving meanings from sentences, in which a sentence is analysed by combining the meaning of parts of that sentence. The distributional hypothesis focuses more on the meaning of an individual word by using a statistical approach. According to this hypothesis, the meaning of a word is derived by the context it appears in and can be represented in a (latent) vector space. The approach of capturing the meaning of words using vector representations is also called *word embedding* (Levy and Goldberg, 2014a).

Different word embedding methods exist that have been embraced by the NLP literature. First of all, Mikolov's et al. (2013) introduced *Word2Vec* and the concept of word embedding in their paper (Levy and Goldberg, 2014c). However, this thesis will apply *Global Vectors* (GloVe) as a word embedding approach to capture similarities between tokens in a corpus, for reasons discussed in the next section.

3.2.1 Global Vectors (GloVe)

Global Vectors (GloVe), introduced by Pennington, Socher and Manning (2014), aims to capture word relationships similar to Word2Vec. But whereas Word2Vec focuses on a context window in order to predict words or context words locally, GloVe applies a global approach. In order to do so, a co-occurrence matrix (X) has to be established first. This is a large two-dimensional matrix that represents term-term frequency of all words within a corpus. Additionally, let X_{ij} be the joint occurrence of words i and j, which is the number of times word i appears in the context of word j. Also, $X_i = \sum_k X_{ik}$, which is the total number of times any word appears within the context of word i. Finally, let $P_{ij} = \frac{X_{ij}}{X_i}$, which is the probability that word j appears in the context of word i.

After establishing X, the next step is to consider the objective function, which is a general function that takes the word embeddings for i, j and k as input and shows co-occurrence of words using word vectors. This objective function is the first step in the process of matrix factorization, i.e. transforming the large term-term frequency matrix into a lower-dimensional matrix that represents word embeddings. Similar as for Word2Vec, GloVe aims to capture simple arithmetic relationships between words in a corpus. However, in doing so GloVe considers co-occurrence probabilities as opposed to individual occurrence. The authors of the paper explain this with an example displayed in table 3.1.

Table 3.1: Co-occurrence probabilities of the target words "ice" and "steam" using context words from a 6 billion token corpus (source: Pennington et al., 2014)).

Probability and ratio	k = solid	k = gas	k = water	k = fashion
$P(k\|ice)$	$1.9 * 10^{-4}$	$6.6 * 10^{-5}$	$3.0 * 10^{-3}$	$1.7 * 10^{-5}$
$P(k\|steam)$	$2.2 * 10^{-5}$	$7.8 * 10^{-4}$	$2.2 * 10^{-3}$	$1.8 * 10^{-5}$
$\frac{P(k\|ice)}{P(k\|steam)}$	8.9	$8.5 * 10^{-2}$	1.36	0.96

In this example, some of the context words are related to either one or both the target words, and some are related to none of them. For words that are either related to both target words ("water") or to none of the target words ("fashion"), the co-occurrence ratio is close to 1. For words related to ice but not steam ("solid"), this ratio should be large. Likewise, for words that are not related to ice but are related to steam ("gas"), this ratio is small.

This example shows that co-occurrence probabilities are able to distinguish between relevant and irrelevant words, where individual probabilities fail to do so. For this reason, the general objective function is formulated as followed:

1) $F(w_i, w_j, \widetilde{w}_k) = \frac{P_{ik}}{P_{jk}}$

For which w_i, w_j and $\widetilde{w}_k$ are the word vectors corresponding to i, j and k. As the authors describe, w and $\widetilde{w}$ perform identically and only differ in the initial starting value used for building the word vectors. In their algorithm, the sum of w and $\widetilde{w}$ will be used to build word vectors, because multiple starting points can help reduce noise and overfitting of the model (Ciresan et al., 2012).

To assure arithmetic possibilities between vectors, F should also include arithmetic's in order for vectors to having meaning. Therefore, F's input should capture the difference between the word vectors:

2) $F(w_i - w_j, \widetilde{w}_k) = \frac{P_{ik}}{P_{jk}}$

In this equation, the left-hand side contains vectors, while the right-hand side is a scalar. To solve this, we can create a linear relation between the word vectors w_i, w_j and $\widetilde{w}_k$. This can be obtained by taking the *dot product* of these word vectors. The dot product assures scalarity by considering the Euclidian magnitude (distance) and the angle of a vector. For vectors a and b it is calculated by multiplying the length of these vectors (which is the product of Euclidian magnitude), multiplied by the cosine of the angle of the vectors.

3) $F(dot(w_i - w_j, \widetilde{w}_k)) = \frac{P_{ik}}{P_{jk}}$

Additionally, it is important that the formula allows interchangeability of context words and other words in the corpus. In order to achieve this, it is necessary to assure homomorphism when changing between context words and other words. This means that we should be able to change the inputs of the model while preserving its natural structure. This is achieved by require F to be a homomorphism:

4) $F(dot(w_i - w_j, \widetilde{w}_k)) = \frac{F(dot(w_i), \widetilde{w}_k)}{F(dot(w_j), \widetilde{w}_k)}$

Solving this by using equation (3) leads to:

5) $F(dot(w_i), \widetilde{w}_k) = P_{ik} = \frac{X_{ik}}{X_i}$

Next, the authors take the log of the probability ratio in order to be able to subtract the probabilities:

6) $dot(w_i), \widetilde{w}_k = log(P_{ik}) = \log\left(\frac{X_{ik}}{X_i}\right) = \log(X_{ik}) - \log(X_i)$

As stated before, we want to be able to switch between context words. Also, we previously defined X_{ik} being equal to the number of times word i appears in the context of word k. Therefore, $X_{ik} = X_{ki}$ because word i only appears in the context of k if word k appears in the context of word i. Equation (6) would thus suggest exchange symmetry if not for the log (X_i) on the right side. To solve these problems, the authors replace X_i with a bias term for word i since X_i is independent of k. Finally, a bias term for word k is included to make the equation symmetric:

7) $dot\big((w_i), \widetilde{w}_k\big) + b_i + \tilde{b}_k = \log(X_{ik})$

This equation forms the basis of the GloVe model. However, the authors recognize that some issues are not being dealt with by this formula. One of them is the occasion in which the number of co-occurrences is equal to 0. Furthermore, they admit that this model weighs all co-occurrences equally, even if they appear very frequent or infrequent. Words that appear very infrequent should be considered noise, whereas words that appear very frequent should not be emphasized. Therefore, these words should not be overweighed by the model. The authors deal with these issues by introducing a cost function which is a least squares regression model based on equation (7). Furthermore, to deal with the very frequent or infrequent words, they introduce a weighting function $f(X_{ij})$. This leads to the following equation:

8) $J = \sum_{i,j=1}^{V} f(X_{ij})(\, dot\left((w_i), \widetilde{w}_j\right) + b_i + \tilde{b}_j - \log(X_{ij}))^2$

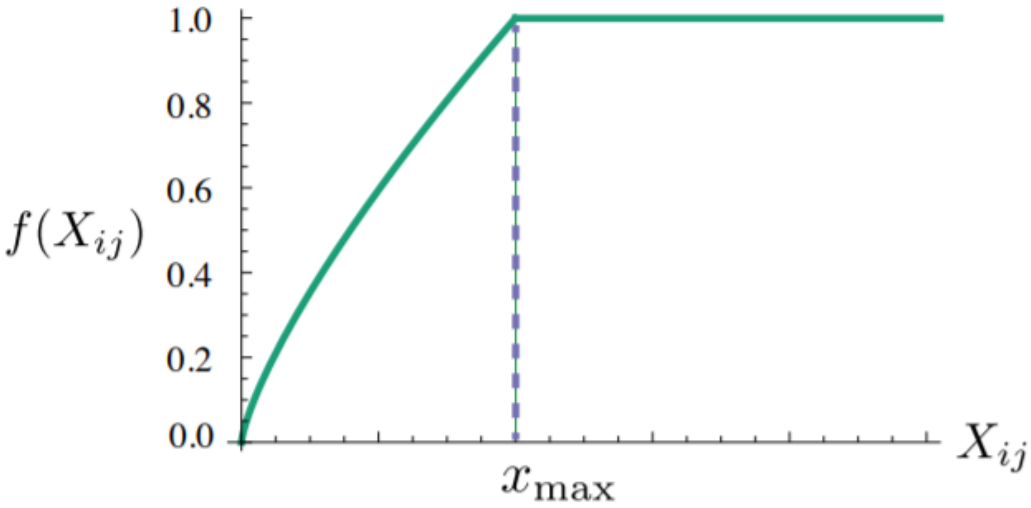
For which V is the size of the vocabulary. Finally, the authors discuss the properties that this weighting function $f(X_{ij})$ should contain in order for it to be meaningful. For $X = 0$, the weighting function should not fall below 0. Additionally, $f(X_{ij})$ should be non-decreasing to prevent infrequent co-

occurrences from being overweighed and $f(X_{ij})$ should be small for very frequent appearing co-occurrences. To meet these requirements, the following weighting function has been formulated:

9) $$f(x) = \begin{cases} \left(\frac{x}{x_{max}}\right)^{\alpha} & if\ x < x_{max} \\ 1 & otherwise \end{cases}$$

For which x_{max} is not equal to the largest number of co-occurrences of a pair of words in a corpus, but is a fixed constant in the cost function (the authors use $x_{max} = 100$). α is an undefined parameter which the authors set at $\frac{3}{4}$ based on empirical testing. As the weighting function displays, for values $x < x_{max}$ the value of $f(x) = 1$. This assures that the weighting of co-occurrences will always lie between 0 and 1. Graphically, the weighting function is displayed in figure 2.2.

Figure 2.2: Weighting function f with $\alpha = \frac{3}{4}$ (source: Pennington et al., 2014).



Whereas the dot product considers the Euclidian magnitude and angle of a vector, cosine similarity only cares about the angle of vectors. Cosine similarity measures the cosine of the angle between two vectors in a (latent) vector space. For two vectors (A and B), cosine similarity is equal to:

$$\cos(A, B) = \frac{A * B}{||A|| * ||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2} \sqrt{\sum_{i=1}^{n} {B_i}^2}} \in [-1, 1]$$

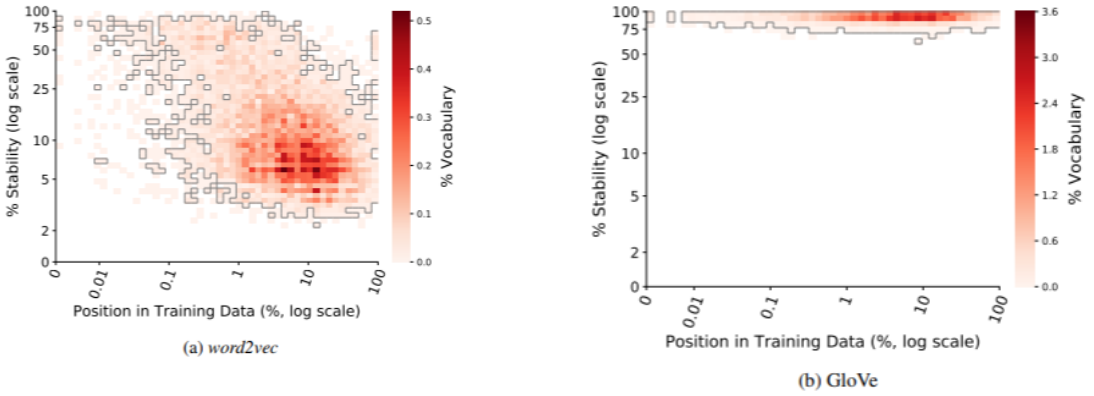
The cosine is equal to 1 for vectors with an angle of 0°, 0 for vectors with an angle of 90° and -1 for vectors with an angle of 180°. Cosine similarity is used in GloVe to determine word similarity because differences in word frequencies (magnitude) should not be relevant in determining word similarity.

Both Word2Vec and GloVe share the objective of producing word embeddings based on co-occurrence of target word and context words, which can be used to observe semantic relationships between

words using simple arithmetic operations. However, Word2Vec and GloVe differ in the way in which these word embeddings are produced. For example, Word2Vec captures word co-occurrence once at a time (using a fixed context window) and produces word embeddings using a three-layered forward-feeding neural network. In doing so, Word2Vec maximizes the average log likelihood of words occurring within a context (*CBOW*) or context words appearing nearby a target word (*Skip-gram*) by applying a softmax classification model. Therefore, Word2Vec is considered a predictive model (Mikolov et al., 2013). In contrast, GloVe's starting point is not a input layer in a three-layered neural network, but a large co-occurrence matrix which captures the count of co-occurrences of all words within a corpus. This matrix represents the term-term frequency of words and is therefore a large two-dimensional matrix. By using matrix factorization, this large matrix get transformed in a matrix in which word embeddings for each word are represented as a row.

Word2Vec and GloVe are compared in completing numerous NLP tasks. Some authors conclude that GloVe performs better at word analogy and word similarity tasks compared to SVD and CBOW (Pennington et al., 2014). However, other authors find the opposite and conclude that Skip-Gram outperforms GloVe on most word analogy and word similarity tasks (Levy et al., 2015). These differences in outcomes between the papers could be explained by both papers using different datasets for their word similarity and word analogy tasks. Therefore, it is difficult to say which algorithm is preferred in building word similarity lexicons. One significant benefit of GloVe compared to Word2Vec however is its stability in producing word embeddings, as described by Wendlandt et al. (2018). In this paper, stability is defined as the percentage of overlap between nearest neighbours in a latent embedding space. In the presence of two different word embedding spaces (for example, X and Y), the authors take the nearest neighbours of word W based on cosine similarity. For both embedding spaces, the overlap for W and its ten nearest neighbours is calculated and the average overlap for all ten neighbours is defined as stability. If X and Y share four similar neighbours of word W, stability of that word is equal to 40%. As figure 2.3 displays, GloVe performs significantly better on stability compared to Word2Vec. Additionally, the authors found that the stability of Word2Vec fluctuates significantly with the frequency of words appearing in a corpus. For words with low frequency, as the number of nearest neighbours increases, the stability of that specific word increases exponentially. For words with a high frequency, choosing a very low or high number of nearest neighbours shows the greatest stability. However, the stability of GloVe remains constant for all word frequencies. For this reason, GloVe will be applied in building the lexicons. The creation of these lexicons will be discussed extensively in section 4.5.

Figure 2.3: Stability of both Word2Vec and GloVe based on the starting position of word W in the training data (source: Wendlandt et al., 2018)



3.3 Multiple linear regression

In multiple linear regression, the relation between a dependent variable and a set of predictor variables is defined as a linear combination displayed in the following model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

For which X_p represent the predictor variables, β_p represent the corresponding variable coefficients and ϵ is an error term. The variable coefficients are estimated and are chosen by minimizing the *sum of squared residuals* (RSS) of the coefficient estimates:

$$\sum_{t=1}^{n} (y_t - \hat{y_t})^2 = \sum_{t=1}^{n} (y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2$$

In order to obtain reliable coefficient estimates, several assumptions have to be met. First, as displayed in the general model, multiple linear regression requires a linear combination between the regression coefficients and the error term. By satisfying this condition, the change in the dependent variable can be assessed as a result of a one-unit increase in X_p, without regarding the value of X_p. The second assumption requires the residual errors to be normally distributed with a mean of 0 and variance σ. Also, the principle of homoscedasticity requires the error term to be constant over all values of X_p. The additive assumption states that the influence of any input variable on the dependent variable is independent of a change in value for any other input variable. Interaction terms can be included in the model if there is any concern that this might not be the case. An interaction term considers the impact of two input variables on the dependent variable simultaneously. Finally, multicollinearity should not be present in the model. Multicollinearity will be further discussed in section 3.4.3. Diagnostics will be run to check whether one of these assumptions are violated. These will be discussed in detail in section 4.2.

Outliers can negatively influence the reliability of the model and corresponding variable coefficients. Cook's distance will be used as a metric to identify outliers (Cook, 1977). Cook's distance (D_i) identifies outliers as a measurement over the residuals and leverage (distance of variable x_i for one observation relatively to all other observations) of each observation. This is denoted in the following formula:

$$D_i = \sum_{j=1}^{n} \frac{(\hat{Y}_j - \hat{Y}_{j(i)})^2}{(p+1)\hat{\sigma}^2}$$

For each observation, Cook's distance measures the impact of deleting that observation on the change of the predicted dependent variable (Kim et al., 2001). A general guideline for labelling observations as outliers is to use a threshold of Cook's distance equal to 1. This threshold will also be used in this thesis. The process of identifying and removing outliers is discussed in section 4.2.

3.3.1 Model performance

Model performance of multiple linear regression will be assessed using *residual standard error* (RSE) and R^2. As the name indicates, RSE is an estimate of the standard error of the residuals ϵ (James et al., 2013). In formula:

$$RSE = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

RSE represents the average deviation of the dependent variable from the true regression line. R^2 resembles the proportion of variance explained by the model. For multiple linear regression, it is equal to the squared correlation of the true- and predicted values of the dependent variable ($Cor(Y, \hat{Y})^2$). Since it represents a proportion, the value of R^2 will always lie between 0 and 1. The closer the R^2 gets to 1, the more variance the model captures of the dependent variable.

3.4 Logistic regression

Logistic regression is a supervised learning method for which the dependent variable is binary. In contrast to multiple linear regression, the prediction given any combination of independent variables will always lie between 0 and 1. The logistic function lies at the foundation of this attribute. The logistic function is as followed:

$$p(x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

This function can be rewritten into the following equation:

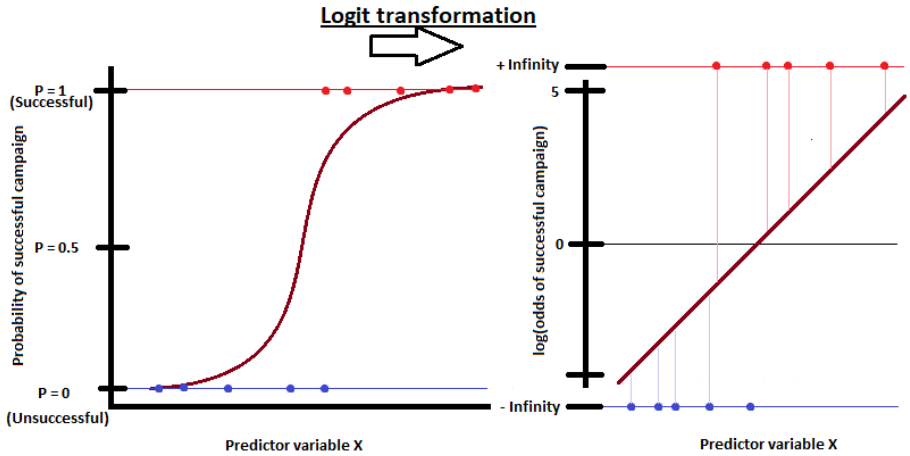
$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 X}$$

The left-hand side of this function is being referred to as *odds* and is simply the probability of an event happening divided by the probability of an event not happening. For example, when the probability of a successful crowdfunding campaign is equal to 0.4, the odds are $\frac{0.4}{1-0.4} = \frac{2}{3}$, so the odds are 3 to 2 that a campaign will not be successful. And by taking the logarithm of both sides:

$$log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 X$$

For which the left-hand side is called the *log odds*. This transformation is convenient, because we can draw a "best fitting" line on the graph (similar as for multiple linear regression). Figure 3.1 shows this transformation graphically.

Figure 3.1: Graphical display of logit transformation.



As this figure shows, the logit transformation converts the raw observations to positive- and negative infinity. Although it is possible to draw a "best-fitting" line based on the transformed observations, finding the exact values for the regression coefficients using least squares is not possible. The transformations to positive- and negative infinity means that the residuals are also equal to positive- and negative infinity. Therefore, estimating coefficients similarly as for multiple linear regression is not possible. To solve this, the regression coefficients can be estimated using maximum likelihood instead (James et al., 2013). In formula:

$$L(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i\prime:y_{i\prime}=1} (1 - p(x_{i\prime}))$$

For which $p(x_i) = \Pr(Y = 1 \mid x_i)$ which is the probability of a successful campaign given variable x_i based on Bernoulli distribution. The goal of this equation is to find values for β_0 and β_1 such that the predicted probabilities are as close to the observed values of the dependent variable (either 0 or 1). This maximizes the likelihood function. Solving for β_0 and β_1 has to be done iteratively, since there is no closed-form solution.

Applying logistic regression as a supervised learning method requires several assumptions. First, observations have to be independent of each other (no repeated measurements). Second, the predictor variables have to be linearly related to the log odds. Finally, multicollinearity amongst predictor variables should be avoided. Multicollinearity will be discussed in section 3.4.3.

Table 3.2: Example of logistic regression output using one predictor variable.

Variable	Coefficient	Std. error	Z-statistic	P-value
Intercept	-10.6513	0.3612	-29.5	<0.0001
Donation amount	0.0055	0.0002	24.9	<0.0001

The output of the logistic regression (table 3.2) can be interpreted using the logistic function. For a regression model with one predictor variable (*donation amount*), the probability of a campaign being successful with $1.000 donated to is equal to:

$$\hat{p}(X) = \frac{e^{-10.6513+0.0055*1000}}{1 + e^{-10.6513+0.0055*1000}} = 0.00576$$

Using a threshold of $\hat{p}(X) > 0.5$ for campaigns to be successful, this campaign will be classified as unsuccessful by the model. Another way of interpreting individual coefficients is by transforming the log-odds into odds. The log-odds can be defined as the expected change for increasing a variable by one unit and is equal to the logistic regression coefficient for that variable. For example, an increase in *donation amount* of 50 units increases the log-odds of campaign success by $50 * 0.0055 = 0.275$. The odds ratio can be computed by exponentiating this coefficient. Thus, an increase of *donation amount* of 50 units increases the probability of campaign success by $e^{50*0.0055} = 1.316$ which is an increase of 31.6% in campaign success.

3.4.1 Model performance

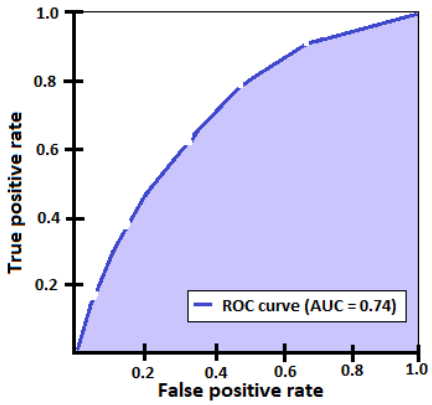
Model performance for different combinations of variables will be compared based on their corresponding Akaike Information Criterion (AIC) estimate. In formula, AIC is computed as followed:

$$AIC = 2k - 2\ln(\hat{L})$$

For which $\hat{L}$ is equal to the maximum likelihood of a model and k the number of variables. Thus, the AIC is dependent of the log-likelihood and penalized by the number of variables used. The model for which AIC is the lowest will be selected as final model for the logistic regression.

Variable importance for the logistic regression will be determined based on the *area under the ROC curve* (AUC). For each predictor variable, the true positive rate (also known as recall, discussed in section 3.4.2) is plotted against the false positive rate ($\frac{FP}{FP+TN}$) by using a series of cut-offs to the predictor data to predict the class ("Variable importance", 2019). An example of AUC is displayed in figure 3.2.

Figure 3.2: Example of ROC curve and corresponding AUC.



3.4.2 Model prediction

Model accuracy for binary classification tasks can be computed by summing the number of correctly predicted clusters ($TP + TN$), divided by the total amount of observations. As can be seen in figure 3.3, TP is the total number of correctly predicted successful campaigns (1) and TN is the total number of correctly predicted unsuccessful campaigns (0). However, for classification tasks in which the predicted classes are unevenly distributed in the data, F1-scores are suggested to provide a better insight in the quality of the model. For unevenly distributed classes, a predictive model can achieve a high accuracy by simply choosing the majority class (Powers, 2011). F1-scores are comprised of precision and recall. Precision is the number of correctly predicted successful campaigns (TP), divided by the total number of predicted positives ($TP + FP$). Recall is the number of correctly predicted successful campaigns (TP), divided by the total number actual positive classes ($TP + FN$). In formula, F1-scores are computed as followed:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

When the costs of predicting false positives is high (for example, predicting survivors of a plane crash), precision is a good measurement because it focuses on number of correctly predicted positives. In a similar way, recall is the correct measurement when the costs of predicting false negatives is high. F1-scores provide a balance between precision and recall because it takes the harmonic average of precision and recall (Powers, 2011). Thus, F1-scores are more appropriate than accuracy when the predicted classes are unevenly distributed in the data and more appropriate than precision and recall when the costs of predicting false negative- and positives are even.

Figure 3.3: Confusion matrix for computing model accuracy and F1-measures for binary classification tasks.

		Predicted clusters	
		0 (Unsuccessful)	1 (Successful)
Actual clusters	0 (Unsuccessful)	True negatives (TN)	False positives (FP)
	1 (Successful)	False negatives (FN)	True positives (TP)

3.4.3 Multicollinearity

The absence of multicollinearity is one of the assumptions of logistic regression and multiple linear regression. High correlation between predictor variables causes their respective regression coefficients to be unstable. For these variables, interpreting the effect of an increase by one unit on the dependent variable holding all other variables constant is not safely assumable because these coefficients are likely to change as a result of small changes in the model. Therefore, it is sensible to check for multicollinearity before selecting variables as input for the model. Multicollinearity can be detected by computing a correlation matrix of all predictor variables considered for the analyses. Using Pearson correlation coefficients, correlation coefficients range between -1 to +1, for which an absolute value of 1 indicates perfect linear relationship between two variables. Correlation coefficients close to this value should not be present in each of the regression models. This thesis uses a correlation coefficient threshold of 0.8 for variables to be considered problematic.

4. Data

Data has been collected from the website of GoFundMe. First, the URLs of campaigns within each category have been scraped. In order to be able to interact with the 'see more' button, *Octoparse V7.1.2*[1] has been used to collect the URLs. These URLs have been imported into R and the data has been collected using the *rvest*-package. Because of different laws with regards to crowdfunding and other cross-country differences such as healthcare insurance policies, only campaigns have been scraped that are established in the United States of America. Furthermore, only campaigns that are active for over a month have been used. This prevents very recent campaigns from skewing the analyses. As a result, the initial 8,335 campaigns used in this analyses are created between 4th of October, 2018 and 4th of April, 2019 and the data has been collected on the 18th of May.

4.1 Cleaning and Variable extraction

4.1.1 General cleaning for numerical– and categorical variables

First, campaigns that have been closed by the initiator have been removed. For these campaigns, GoFundMe only provides a fraction of the information compared to active campaigns. Additionally, campaigns that have unrealistically high goal amounts will be removed. Whereas Kickstarter applies an 'all-or-nothing' approach in which the campaign creator only receives the funds collected when the set goal amount is reached, GoFundMe does not apply such a donation structure. The campaign creator can therefore set his goal amount to whatever he pleases, without any consequences. This thesis will use a similar approach as Mollick (2014) and Pitschner and Pitschner-Finn (2014) to set a boundary for donation goals to be considered unrealistic. Accordingly, campaigns with a donation goal below $100 and above $10,000,000 are removed. Finally, campaign descriptions with less than 5 words used are removed. All numerical variables have been directly scraped or are indirectly derived from the GoFundMe campaigns and are based on findings of successful campaign success predictors described in other literature, which is discussed in section 2.3. In this thesis, success rate will be used as the dependent variable for the logistic regression. A campaign is classified as successful if the donation amount exceeds the goal amount. Additionally, the number of Facebook shares of the campaign will be used as an indicator of WOM. Furthermore, sentiment score of a campaign has been included as a numerical feature. Sentiment score will be discussed in more detail in section 4.4. All numerical variables have been standardized. The categorical variables have been converted into dummies using one-hot encoding.

[1] Available at: https://www.octoparse.com/download

4.1.2 Descriptive statistics for numerical- and categorical variables

An overview of the numerical - and categorical variables can be found in tables 4.1-4.4 and figure 4.1.

Table 4.1: Overview of numerical variables.

Variable	Description
Donation amount	Number of dollars donated to a campaign.
Goal	Donation goal set by the initiator.
Number of donators	Number of donators to a campaign.
Number of Facebook shares	Number of campaign shares on Facebook.
Average donation	Average number of dollars donated to a campaign.
Number of words	Number of words used in the campaign description.
Sentiment score	Sentiment score of a document based on the Bing lexicon.
Topic 1-20 (LDA)	Probability of a document belonging to a specific topic.
Total numerical variables	27

Table 4.2: Descriptive statistics of numerical variables (before standardizing).

	Donation amount	Goal	Donators	Shares	Average donation	Word count	Sentiment score
Mean	$16,978	$31,720	179	708	$132	316	5
(SD)	($31,975)	($108,651)	(395)	(2,345)	($268)	(394)	(6)
Median	$5,190	$10,000	62	206	$91	230	4
[Q1, Q3]	[$2,770, $22,679]	[$5,000, $30,000]	[30, 188]	[54, 639]	[$63, $137]	[133, 388]	[1, 8]
Min - Max	$715 - $1,330,290	$1,000 - $6,000,000	1 - 17,941	0 – 114,000	$13-$15,000	1 – 22,471	(-47) - 56

Table 4.3: Overview of categorical variables.

Variable	Description
Category	Category in which the campaign is listed by the campaign initiator (categorical, 14 in total).
NPO	Binary, “1” if the campaign is initiated by a NPO, “0” if otherwise.
Success	Binary, “1” if the campaign goal has been reached, “0” if otherwise.
Active	Categorical, describes how many months the campaign has been active from date of creation until date of data collection (Short = less than two months, Medium = longer than two months and less than four months, Long = longer than four months).
Reciprocity	Binary, “1” if the campaign contains at least one n-gram from the reciprocity lexicon, “0”otherwise.
Altruism	Binary, “1” if the campaign contains at least one n-gram from the altruism lexicon, “0”otherwise.
Total categorical variables	**6**

Figure 4.1: Category – occurrence and frequency.

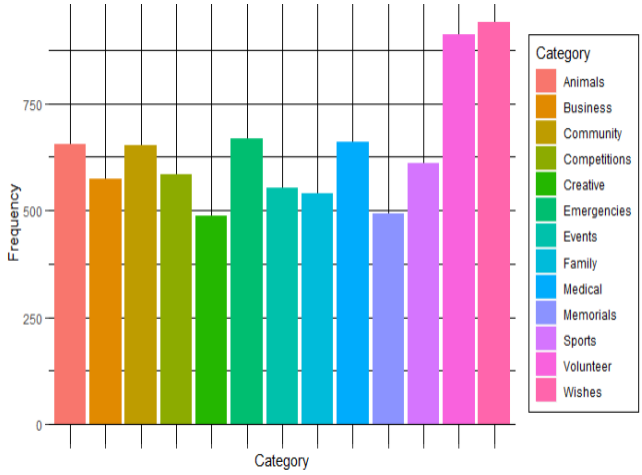


Table 4.4: Descriptive statistics of categorical variables.

Variable	N	Share of campaigns
NPO	1,869	22.5%
Success	2,397	28.9%
Reciprocity	639	7.7%
Altruism	1,830	22.0%
Active_short	1,442	17.4%
Active_medium	2,726	32.8%
Active_long	3,631	43.7%

4.1.2 General cleaning for semantical variables

General data cleaning has been applied to prepare the campaign description for the semantical analyses. This includes removing excess spaces in the text, removing abbreviations and replacing them with the full corresponding terms, removing numbers, removing capital letters and replacing them with lowercase letters and removing special characters. Furthermore, stop words are removed. The semantical variable extraction will be discussed in the Section 4.3 and 4.4. Section 4.3 will discuss the extraction of topics from the corpus. Section 4.4 focuses on assigning sentiment scores to documents.

Figure 4.2: Top 25 most frequent appearing words in the campaign description.

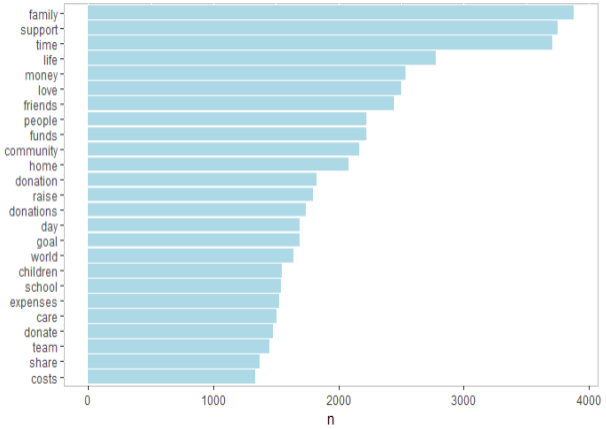


Figure 4.3: Word cloud based on the 200 most frequent occurring words in the campaign description.

4.2 Multiple linear regression diagnostics

Different diagnostics are run to check if the data does not violate the assumptions of multiple linear regression. First, we will check if the dependent variable (*donation amount*) is normally distributed. The corresponding histogram is displayed in figure 4.4.

Figure 4.4: Histogram of the distribution of donation amount.

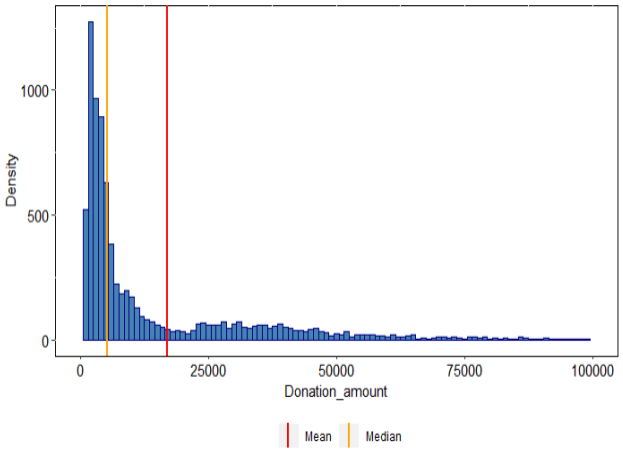
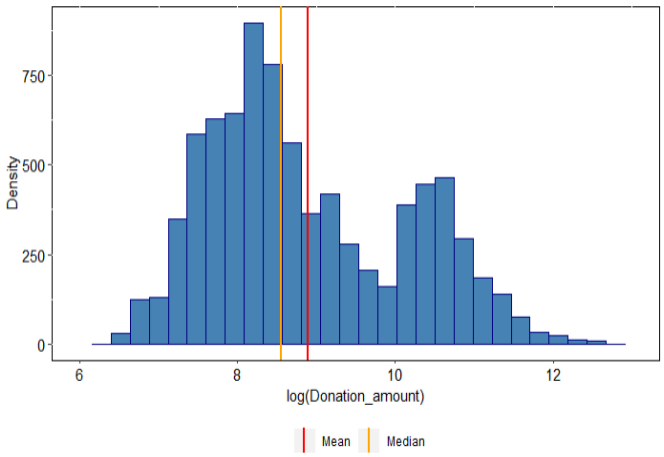


Figure 4.5: Histogram of the distribution of log(donation amount).



The histogram on the left shows that *donation amount* is right-skewed. *Donation amount* as a dependent variable will therefore be log-transformed. This allows for the dependent variable to be normally distributed (figure 4.5). Next, we check whether the linearity assumption or the normal distribution of residuals is violated. These plots are shown in figure 4.6 and 4.7 respectively.

Figure 4.6: Residual plot for linearity check.

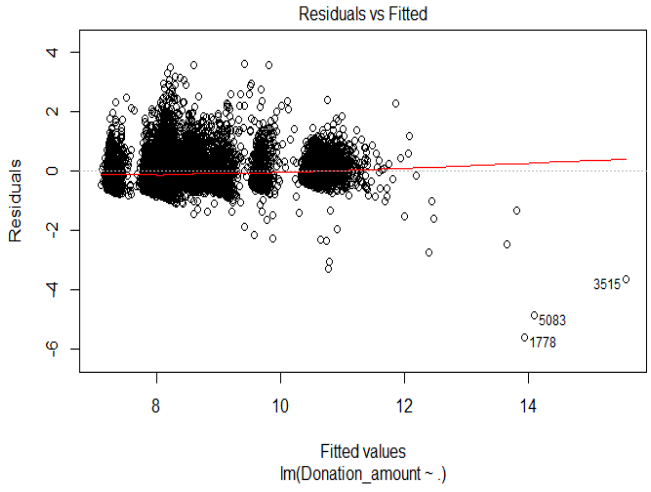
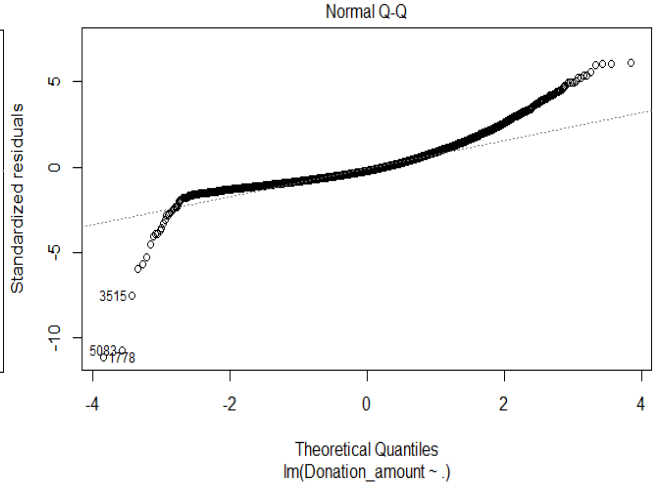


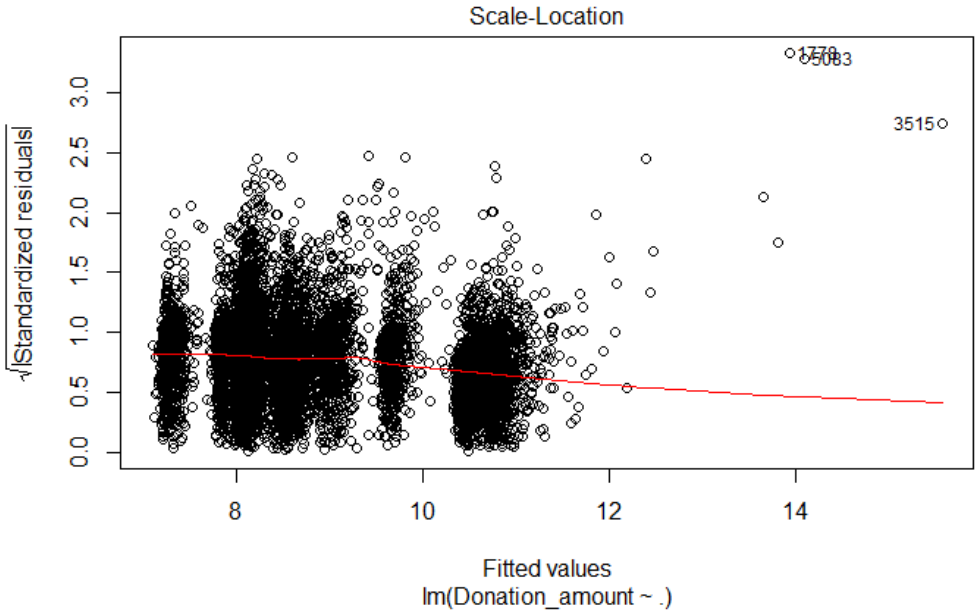
Figure 4.7: Normality of residuals plot.



To accept the linearity assumption, the fitted line in figure 4.6 should be horizontal and approximately at zero. Therefore, we can safely accept the linearity assumption. The normal distribution of residuals is accepted when the residuals follow a straight line. This is the case according to figure 4.7, so we can

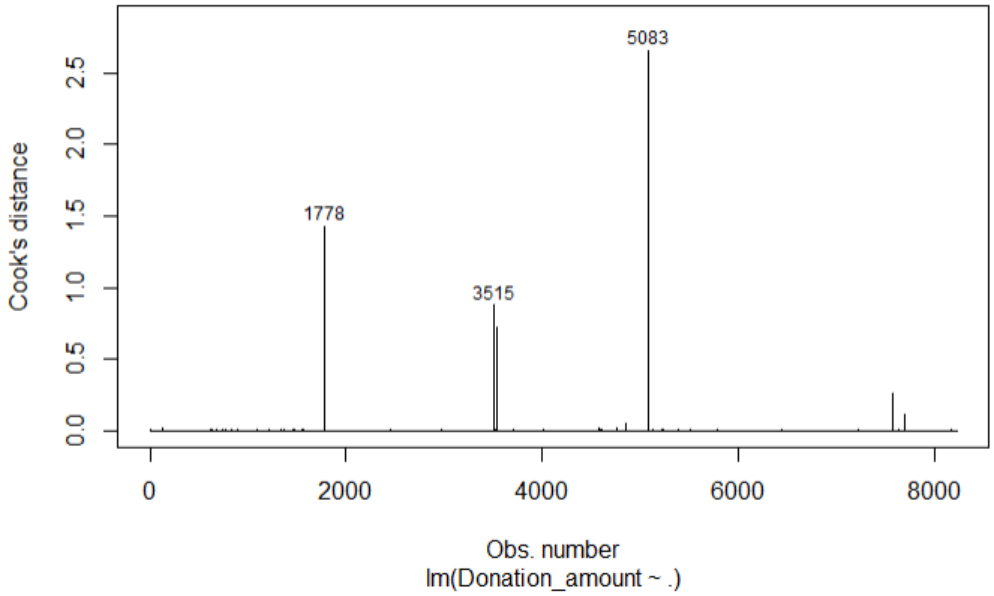
assume a normal distribution of the residuals. Homoscedasticity can be checked by plotting the standardized residuals for the different outcome values of the dependent variable, for which a straight line indicates constant variability of the residuals. Possible solutions of heteroscedasticity include log-transforming the dependent variable, which already has been done previously. As figure 4.8 shows, we can safely assume the absence of heteroscedasticity.

Figure 4.8: Variability of residuals plot.



Some concerns have to be mentioned with the previous figures. Outliers are visible in all three plots, which might have a negative impact on model performance. To assess the impact of these outliers on model performance, we will run multiple linear regression for the models with- and without outliers and compare RSE for both models. Cook's distance will be used as a metric to identify outliers. Cook's distance (D_i) identifies outliers as a measurement over. Figure 4.9 shows the Cook's distance for all observations graphically. These observations will be removed from the data manually and RSE will be compared. RSE for the full model is equal to 0.5875. Four observations have been removed from data which have been labelled as outliers. RSE for the model without outliers is equal to 0.5741. The model without outliers will therefore be used for the multiple linear regression.

Figure 4.9: Cook's distance plot.



4.3 Latent Dirichlet Allocation

As explained in section 3.2, two hyperparameters have to be determined in order for topic distributions to be assigned to the campaign descriptions. The first hyperparameter is the number of topics (k). The stemmed campaign descriptions are transformed into a *document-term matrix* (DTM), in which the rows represent the campaign descriptions (8,231 in total) and the columns represent the unique tokens included in the campaign descriptions (24,019 in total). To filter the amount of tokens considered for each topic, token importance based on their corresponding TF-IDF has been computed and only tokens for which this score is greater than 0.1 are considered in determining the topics. LDA has been performed on these tokens iteratively for k between 5 and 60 on the training data, which consists of 80% of the documents, and for each value of k been tested on a validation set. The results of this tuning process are displayed in appendix A, figure 8.1. As can be seen in this figure, the perplexity for the validation set is lowest at $k = 20$, which is the number of topics considered for the rest of the analysis. Next, tuning parameter (α) has been set based on k and has been run iteratively for α between 0.01 and 0.2 on the training data and been tested on the validation set for each value of α. The results of this are displayed in appendix A, figure 8.2. As can be seen in this figure, perplexity for α in the validation set is at a minimum at $\alpha = 0.03$.

Using these parameters, LDA has been applied on the DTM. The topics gained as a result will be used as control variables. As explained in section 2.5.1, these topics can provide other semantical insights by interpreting those topics that are significant predictors of campaign success or total amount funded. The top-50 tokens per topic are used to coherently label each topic. Topics for which a coherent label

cannot be defined or topics that have an overlap in theme will be removed. Additionally, irrelevant tokens included in the topics are dismissed. These tokens are mostly made up from the (common) names of the persons for which the campaign is set up. LDA assigns topic distributions to all documents in the corpus. Thus, for all twenty topics extracted from the corpus a probability of a document belonging to a specific topic is computed. These per-document topic distributions sum to one per document for all topics. These will be transformed into predictor variables by applying one-hot encoding. For each document, topics with a per-document topic distribution greater than 0.7 will be assigned a "1" for the corresponding topic and "0"s for all remaining topics for that document. This ensures that documents will not erroneously be assigned a (small) value for topics due to noise included in that topic. Out of the twenty topics, six could be coherently labelled based on the tokens included in them. These six topics including the corresponding top-10 token distribution and label are displayed in appendix A, figure 8.3. Table 4.5 displays the descriptive statistics of these topics.

Table 4.5: Descriptive statistics of topics obtained by LDA.

Topic	**N**	**Share of campaigns**	**Example of tokens included**
Topic 5 (*Crime-related*)	175	2.1%	*Stolen, robbed, murdered, conviction*
Topic 8 (*Natural disasters*)	197	2.3%	*Wildfire, burn, flood, belongings*
Topic 11 (*Students*)	150	1.8%	*Pageant, sorority, contestant, teen*
Topic 13 (*Sports competition*)	326	4.0%	*Championship, tournament, Olympics*
Topic 15 (*Cancer victims*)	156	1.9%	*Leukaemia, myeloid, glioma, Disney*
Topic 20 (*Hispanic language*)	106	1.3%	*Nuestro, mayo, pido, sal, paso*

4.4 Sentiment score

As discussed in section 2.6.1, sentiment analysis can be applied to capture differences in polarity used in campaign descriptions within the dataset used in this thesis. This thesis will use the NRC lexicon to compute a sentiment score for each review. Similar as for other sentiment lexicons, NRC assigns a "-1" for each term containing a negative sentiment and "+1" for terms containing a positive sentiment. The sentiment score per description is therefore equal to the sum of the sentiment scores per term used in the description. These sentiment scores are normalized and included as predictor variable for the logistic regression. Additionally, the normalized number of words used in each campaign description will be included as a control variable. These two variables can be found in tables 4.1 and 4.2.

4.5 Lexicon creation

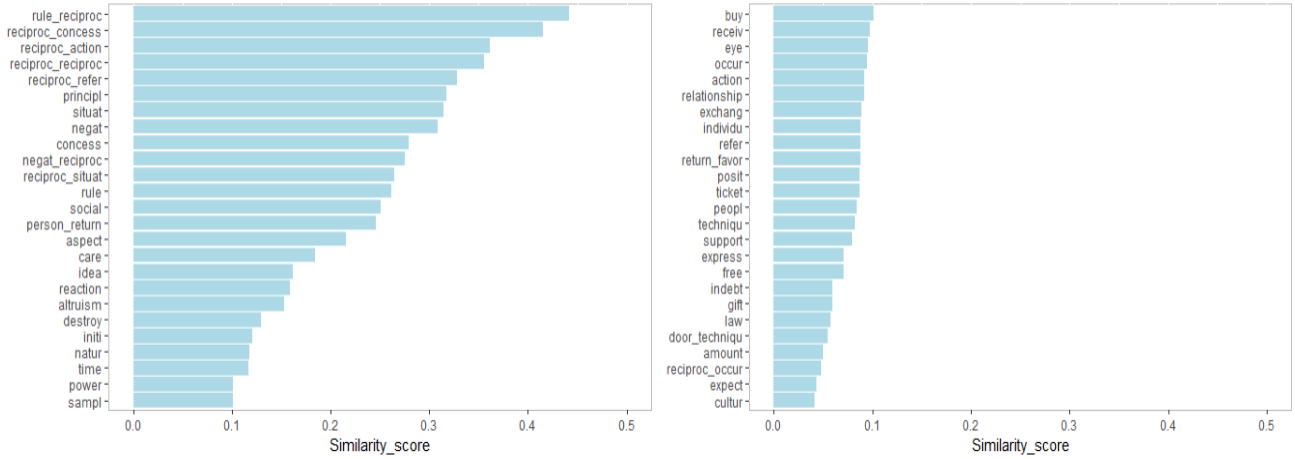
As described in the literature review, concept-specific lexicons will be created which includes similarity scores for both reciprocity and altruism. To accomplish this, the GloVe algorithm will be applied and

similarity scores will be computed based on cosine similarity. Lexicons will be build using online dictionaries and articles that discuss reciprocity and altruism in a crowdfunding perspective. These corpora are manually collected and pre-processed in R. The foundation of the lexicons consists of the unigrams and bigrams that have the highest similarity scores to the tokens “reciproc” and “altruis”. Section 4.5.1 will discuss the reciprocity lexicon and section 4.5.2 the altruism lexicon.

4.5.1 Reciprocity lexicon

Different kinds of articles have been used to create the reciprocity lexicon. These are articles defining and describing the principle of reciprocity in a general manner and articles that focus on reciprocal rewards used in crowdfunding and their benefits. Using these corpora, the top 50 n-grams with the highest similarity to “reciproc” have been included in the lexicon. These are displayed in figure 4.10.

Figure 4.10: Top 50 stemmed n-grams with highest similarity to ‘reciproc’.



As can be seen in this figure, further cleaning of the lexicons have to be applied in order to end up with n-grams that are similar to “reciproc”. For example, “principl” has a high similarity score to “reciproc”, but only because reciprocity is sometimes being referred to as the “principle of reciprocity” and not because “principle” can be used as a synonym for “reciprocity”. The same applies for tokens like “negat” and “posit” (positive- and negative reciprocity) and “rule” and “idea”. These words have therefore been removed from the lexicon. Furthermore, upon inspection of the random subsample of 100 reviews that contain tokens similar to “reciproc”, reviews were erroneously classified as including reciprocal cues because words similar to “reciproc” are used in multiple occasions. For example, “gift” could potentially indicate reciprocal donation intentions when used in a setting that describes campaign initiator rewarding donators by sending them a small gift. However, “gift” can also be used in a setting in which the birth of a child or some other exciting moment in a person’s life is being referred to as “a gift from God”. In that case, “gift” has nothing to do with alluding to reciprocal

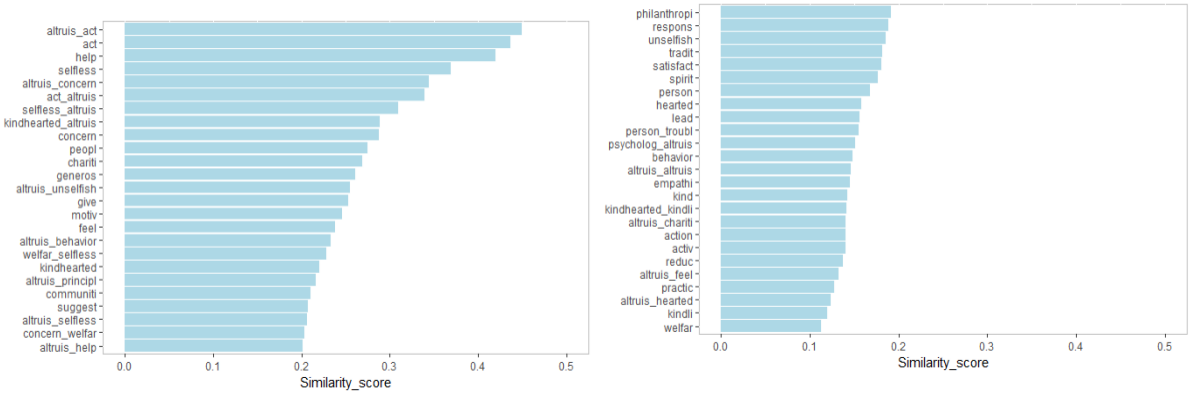
donation intentions. In order to separate these instances, it is important to look at the context at which these words are being used. Therefore, “gift” as an unigram has been excluded from the lexicon, but “thank you gift” has been manually added to the lexicon because it indicates reciprocal intentions of the campaign initiator. Furthermore, campaigns that include the usage of “gift” and similar words in combination with phrases like “you will receive” or “you will get” are also assigned to be alluding to reciprocal donation intentions.

By using GloVe as a foundation for the lexicons and manually adjusting words with a high similarity score so they fit more within a crowdfunding setting, campaigns that contain one or more n-grams included in the reciprocity lexicon have been coded “1” and “0” otherwise. Using this methodology, 639 campaigns have been identified as including reciprocal cues. To further inspect the correctness of the tags that have been assigned to a campaign, a random sample of 100 campaigns that are assigned a reciprocal tag have been manually checked for reciprocal donation intentions. Using this method, 74 out of 100 campaigns were correctly given the reciprocal tag.

4.5.2 Altruism lexicon

Reciprocal campaigns are those campaigns in which some sort of benefit of donating is listed. However, as the benefits of donating from an altruistic perspective are only experienced by the donator and are not determined by the initiator, campaign descriptions do in most of the cases not describe these intrinsic benefits a donator receives. For this reason, identifying altruistic donation incentives is much harder than for reciprocal incentives. Some could even argue that campaigns that do not list reciprocal incentives for donating by default aim to collect funds by triggering altruistic motivations of the (potential) donator. Other motives for donating, such as identification with an organization, would however be neglected in that way. The altruism lexicon will therefore focus on identifying the psychological gains a donator receives, listed by the campaign initiator. To achieve this, articles describing altruism and its benefits will be used to build the initial lexicon. The top 50 stemmed n-grams based on these corpora are displayed in figure 4.11.

Figure 4.11: Top 50 stemmed n-grams with highest similarity to 'altruis'.



Similar as for the reciprocity lexicon, cleaning has to be applied to remove tokens that do not necessarily describe altruism. The next step is to make sure that tokens included in the lexicon are solely indicating altruistic donation intentions. For example, using the token "selfless" could indicate altruistic donation intentions when describing the act of donating. This token could however also be used to describe the person for which the campaign is set up. For this reason, "selfless" should only be considered altruistic when accompanied by tokens such as "donation" or "act of giving". The same applies for tokens similar to "selfless", such as "generos" and "unselfish". Campaigns that contain one or more n-grams included in the altruism lexicon are labelled with "1" and "0" otherwise. In total, 1.830 campaigns have been labelled as alluding to altruistic donation intentions.

4.5.3 Non-profit organizations

Campaigns initiated by non-profit organizations will be identified using a similar approach as for the reciprocity and altruism lexicon. GoFundMe includes an option for campaign initiators to get their campaign certified by the corresponding NPO. Campaigns that include such a certification will be used as a foundation for campaigns to be classified as being initiated by a non-profit organization. However, not all campaigns for which funds are collected for a specific NPO use this certification in their campaign. Therefore, a lexicon is built to detect if the funds of a campaign are collected in name of a NPO. As opposed to the altruism and reciprocity lexicon, we are only interested in synonyms of this term. Distance to a token such as "npo" (which is used as a metric of similarity for the altruism and reciprocity lexicon) will therefore not be applied. Instead, word frequency within documents that include synonyms for tokens such as "npo" will be used to build the NPO lexicon. These synonyms are collected from online dictionaries and pre-processed in R. Campaigns that include one of the synonyms included in the lexicon are assigned a "1" and "0" otherwise. By using this method, 1,869 campaigns have been classified as campaigns initiated in name of a NPO. Table 4.6 displays all combinations of campaigns being classified as NPO, containing altruistic and containing reciprocal cues.

Table 4.6: All possible combinations of campaigns for the variables NPO, reciprocity and altruism.

Campaign classification	N	Share of campaigns
Non-profit organization (NPO)	1,869	22.7%
Altruistic cues	1,830	22.3%
Reciprocal cues	639	7.7%
NPO & Altruistic cues	687	8.3%
NPO & Reciprocal cues	204	2.5%
Altruistic & Reciprocal cues	204	2.5%
NPO & Altruistic & Reciprocity cues	89	1.1%

Campaign classification	N	Share of campaigns
Non-NPO	6,362	77.3%
Non-NPO & Altruistic cues	1,143	13.9%
Non-NPO & Reciprocal cues	435	5.3%
Non-NPO & Altruistic & Reciprocity cues	115	1.4%

5. Results

The following section will include the results for the logistic regression and multiple linear regression. First, control variables will be selected by comparing different models and their corresponding AIC. This will be discussed in section 5.1. The variables selected in this section will form the initial model. The model performance based on accuracy and F1-score and general variable importance will be discussed in section 5.2. Sections 5.3 and 5.4 will discuss the logistic regression output for the main variables and the topics. The model performance of multiple linear regression will be discussed in section 5.5. Similarly as for logistic regression, section 5.6 and 5.7 will discuss the multiple linear regression output for the main variables and topics respectively.

5.1 Variable selection – Logistic regression

Initial variable selection for the logistic regression will be applied based on multicollinearity. Correlation plots will be used to visualize the correlation matrix and detect possible presence of multicollinearity amongst predictor variables. These correlation plots can be found in appendix B. Figure 8.4 shows possibilities of high correlation amongst predictor variables around the categories for the categorical variables *active* and the numerical variables d*onation amount*, *goal*, *donators* and *shares*. Using a correlation coefficient of 0.8 as threshold, figure 8.5 shows no reason to remove *active* as a predictor variable. However, high correlation is observed for *donation amount* and *donators.* To control for the size of the campaign, only one of *donation amount*, *donators* or *goal* have to be included in the model. Therefore, initial variable selection is conducted by comparing the performance of models with different combinations of input variables. These results are displayed in table 5.1.

Table 5.1: Results of different combinations of predictor variables used in logistic regression.

Model	Predictor variables included	AIC	Prediction accuracy	F1 score	Most important variables (Based on AUC)
(1)	All - topics	92.35	99.40%	0.9892	goal, donators and donation amount.
(2)	(1) - donation amount - goal	6,622.7	72.71%	0.1420	donators, word count and sentiment score.
(3)	(2) - donation amount - donators	5,705.8	76.24%	0.4041	goal, word count and sentiment score.
(4)	(3) - goal - donators	6,612.4	72.79%	0.1429	donation amount, word count and sentiment score.

The full model shows almost perfect prediction accuracy. The reason for this is that s*uccess* as a dependent variable is determined based on *donation amount* and *goal*. As can be seen in table 5.1, removing *goal* from the set of predictor variables significantly decreases prediction performance of

logistic regression. The importance plot (figure 5.1) shows that *goal* is by far the most important predictor of all variables. Therefore, model 3 will be used for the logistic regression.

5.2 Model performance – Logistic regression

Logistic regression using the initial model has a classification accuracy of 76.24%. The predictions of this model versus the actual values can be found in table 5.2. Furthermore, the performance metrics can be found in table 5.3. The regression output can be found in appendix table 5.4.

Table 5.2: Confusion matrix including classification predictions for logistic regression.

		Predicted clusters	
		0 (Unsuccessful)	1 (Successful)
Actual clusters	0 (Unsuccessful)	1.684	68
	1 (Successful)	519	199

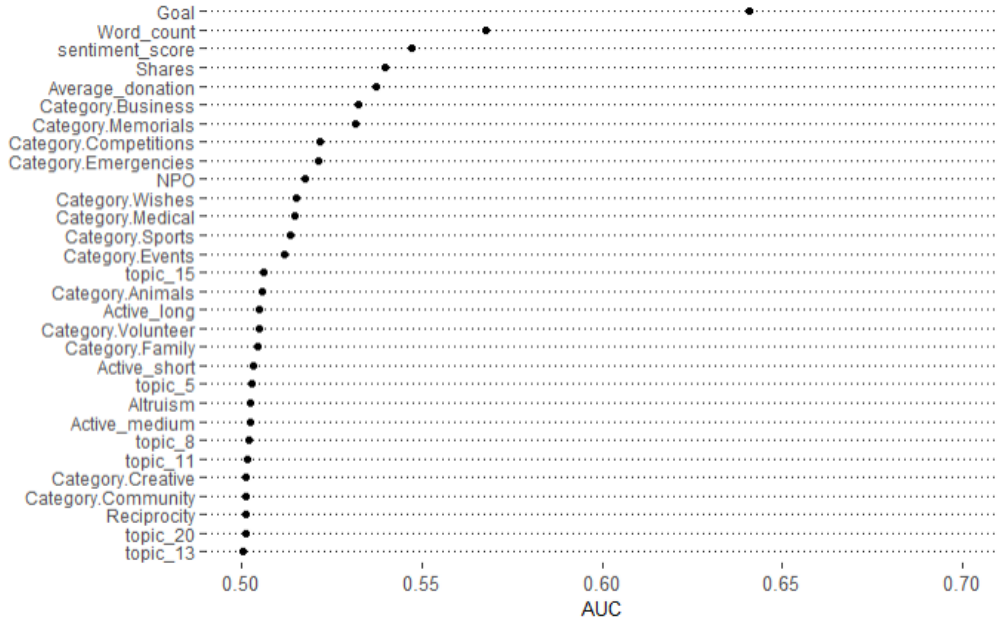
Table 5.3: Performance metrics of logistic regression.

	Accuracy	**Precision**	**Recall**	**F1-score**
Logistic regression	76.24%	74.53%	27.71%%	0.4041

Although a prediction accuracy of 76.24% seems decent, some shortcomings of this model have to be mentioned. First, the total number of unsuccessful campaigns in the dataset (training- and test-set combined) is equal to 71.10% (as can be seen in table 4.4). Thus, classifying all campaigns in the dataset as unsuccessful will result in a prediction accuracy of 71.10%, which is a small decline over the logistic regression prediction. Table 5.3 shows an excellent insight of the shortcomings of this model. Whereas the predictions of unsuccessful campaigns are fairly accurate (1,684 campaigns are correctly classified as unsuccessful and only 68 unsuccessful campaigns are misclassified as successful campaigns), the model struggles to predict successful crowdfunding campaigns. In total, the model predicted only 267 campaigns to be successful, from which 199 are correctly predicted. The poor recall of this model validates these shortcomings described before. Differences in variable selection and distribution between training- and test-set did not significantly improve prediction accuracy. Because successful- and unsuccessful campaigns are unevenly distributed within the data, F1-scores provide more appropriate insights in the quality of the model. The F1-score of this model is equal to 0.4041. As a weighted average of precision and recall, the poor score on this metric is due to the poor recall of the model.

Variable importance (as explained in section 3.4) for logistic regression will be determined by the AUC for each variable. This is displayed in figure 5.1.

Figure 5.1: Overall variable importance plot based on AUC for each individual variable.



Out of all predictor variables, the numeric variables seem to be most important in predicting campaign success, with *goal* being the most important predictor variable. As for the campaign categories, *memorials, business* and *competitions* are the most important predictors of campaign success.

5.3 Main variables – Logistic regression

The logistic regression output (model (1)) can be found in table 5.4. This model will be used to answer hypotheses 1 and 2. The first main variable, *altruism,* is one of the least important predictors of campaign success according to the variable importance plot. The output of the regression shows that *altruism* is insignificant as a predictor variable for the full model. **H1a** is therefore rejected. However, *reciprocity* is a significant predictor (p = .045 using p < .05) of campaign success for the initial model. As *reciprocity* is an unstandardized dummy variable, its effect on campaign success can be assessed by interpreting its coefficient. Using the log-odds function, we can conclude using reciprocal cues in campaign descriptions increases the probability of a campaign being successful by $e^{0.238281} = 1.2691$, which is an increase of 26.91% compared to campaign that do not allude to reciprocal donation intentions (keeping all other variables constant). The significant and positive coefficient of *reciprocity* therefore provides sufficient support for **H2a**. *Sentiment score* is insignificant in predicting campaign success, which means there is a lack of evidence to support **H5a**.

5.4 Topic variables – Logistic regression

To extend the investigation of the effect of semantical features used in campaign description on campaign success, the topics discussed in section 4.3 will be included in model (1). According to the output, topics 5, 11 and 15 significantly influence campaign success. The coefficient of topic 5 is

positive, which indicates that crowdfunding campaigns created for crime-related accidents increase the likelihood of the campaign being successful (keeping all other variables constant). Also, the coefficient of topic 11 is positive. This indicates that crowdfunding campaigns being set up by students are more likely to be successful than campaigns that are not set up by students. Finally, the positive coefficient of topic 15 indicates that campaigns created to support cancer victims are more likely to be successful than campaigns that are not created to support cancer victims.

5.5 Model performance – Multiple linear regression

The results for multiple linear regression to assess the impact of variables on donation amount are displayed in table 5.4, model (2). Multiple linear regression will use the same combination of input variables as those used for logistic regression. R^2 of this model is equal to 0.790 and RSE is equal to 0.577. This indicates that independent variables used as input capture almost 80% of the variance of *donation amount*.

5.6 Main variables – Multiple linear regression

Next, the impact of the variables of interest on donation amount will be discussed. The output of the regression shows that *altruism* is a significant predictor of donation amount (p = .005 using p < .05). These findings support **H1b**. However, the data did not provide enough evidence of the use of reciprocal cues (*reciprocity*) and its effect on total funded amount. Therefore, **H2b** is rejected. The coefficient for the interaction term *reciprocity*NPO* is also insignificant. **H3** is therefore also rejected. The coefficient of *shares* is positive and significant. Using these results, **H4** is accepted. Finally, the coefficient for *sentiment score* is also positive and significant. Before assessing the impact of this variable on the total amount funded, some considerations have to be made. *Sentiment score* is a standardized numeric variable which can take negative values for campaigns containing an overall negative polarity. A positive coefficient indicates that increasing the mean value of *sentiment score* by a certain fraction of its standard deviation increases the total amount funded. Since the mean of *sentiment score* is positive (see table 4.2), the data provides enough evidence to support **H5**.

5.7 Topic variables – Multiple linear regression

Similar as in for logistic regression, additional semantic insights will be explored by including interpretable topics in the multiple linear regression (model (2)) and are also displayed in table 5.4. Topic 8 and 20 turn out to be significant. The coefficient of topic 8 is positive, which indicates that campaigns that are set up to help victims of natural disasters have a positive influence on the total

amount they receive in funds. The negative coefficient of topic 20 indicates that a campaign containing Hispanic language has a negative influence on the total amount funded.

Table 5.4: Regression output for the logistic- and linear regression.

	Dependent variable:	
	Success	ln(Donation amount)
	Logistic regression	*Multiple linear regression*
Model	(1)	(2)
Main variables		
Altruism	0.105 (0.083)	0.055*** (0.019)
Reciprocity	0.249** (0.124)	0.008 (0.030)
Shares		0.268*** (0.012)
Sentiment score	-0.006 (0.041)	0.035*** (0.009)
Interaction variable		
Reciprocity1:NPO1		-0.056 (0.062)
Control variables		
Goal	-6.971***(0.365)	0.109*** (0.009)
Average donation	-0.082 (0.069)	0.092*** (0.010)
Word_count	-0.175*** (0.058)	0.037*** (0.010)
Category Animals	0.146 (0.151)	0.501*** (0.036)
Category Business	-0.689*** (0.194)	-0.042 (0.037)
Category Community	0.711*** (0.148)	0.873*** (0.036)
Category Competitions	-0.832*** (0.174)	-0.784*** (0.038)
Category Creative	0.370** (0.163)	0.431*** (0.040)
Category Emergencies	2.652*** (0.183)	2.203*** (0.037)
Category Events	0.385*** (0.147)	-0.209*** (0.037)
Category Family	1.159*** (0.164)	1.507*** (0.039)
Category Medical	3.544*** (0.210)	2.497*** (0.037)
Category Memorials	2.994*** (0.200)	2.259*** (0.041)
Category Sports	-0.011 (0.156)	0.377*** (0.038)
Category Volunteer	0.278** (0.136)	-0.017 (0.033)
Active(short)	-0.229 (0.155)	-0.049 (0.036)
Active(medium)	-0.275* (0.144)	-0.148*** (0.034)
Active(long)	-0.326** (0.142)	-0.143*** (0.033)
NPO	0.045 (0.086)	0.112*** (0.020)
Topics		
Topic 5	0.398** (0.197)	-0.060 (0.052)
Topic 8	0.183 (0.231)	0.111** (0.051)
Topic 11	0.475** (0.237)	-0.018 (0.059)
Topic 13	0.123 (0.268)	-0.003 (0.042)
Topic 15	0.369* (0.211)	0.083 (0.057)
Topic 20	-0.159 (0.329)	-0.121* (0.067)
Constant	-2.414*** (0.179)	8.282*** (0.038)
Observations	5,761	5,757
R^2		0.790
Adjusted R^2		0.788
Log Likelihood	-2,822.890	
Akaike Inf. Crit.	5,705.779	
Residual Std. Error		0.577 (df = 5727)
F Statistic		741.321*** (df = 29; 5727)
Note:		*p<0.1**p<0.05***p<0.01

6. Conclusion

The aim of this thesis is to identify significant crowdfunding campaign success drivers using a semantical approach. Accordingly, the following research question has been formulated:

What are success drivers of crowdfunding campaigns?

In total, 8,231 campaigns founded in the United States of America from the popular crowdfunding platform *GoFundMe* have been considered for the analyses. The thesis has been split up into hypotheses that assess the impact of a set of variables on crowdfunding campaign success and total funded amount.

H1a, H1b, H2a and H2b are formulated according to theory that describes what drives individuals to donate to a crowdfunding campaign. Sections 2.1 and 2.2 elaborate on altruism and reciprocity being drivers of donation intention. Using campaign descriptions, crowdfunding campaigns are investigated for using language that allude to altruistic or reciprocal donation intentions. Using GloVe as a word embedding method, concept-specific lexicons are built that indicate whether campaign descriptions (as formulated by the campaign initiator) include altruistic or reciprocal donation cues. Altruistic and reciprocal cues are translated to input variables by assigning a "1" to campaigns that include tokens included in the corresponding lexicon and "0" otherwise. Logistic regression is used as a classification method for predicting campaign success and multiple linear regression for assessing the impact of the main variables on the total amount funded. The output of logistic regression indicate that *altruism* is an insignificant predictor of campaign success (using $p < .05$). Therefore, **H1a** is rejected. However, in support with **H1b**, altruistic cues are found to significantly and positively impact total amount funded. This implies that campaigns that utilize altruistic cues in their campaign description received more funds. Additionally, *Reciprocity* turns out to be a significant predictor of campaign success, as opposed to *altruism*. Using reciprocal cues in crowdfunding campaigns (such as rewarding donations with gifts based on the donation amount) therefore does significantly (and positively) influences campaign success, which means that **H2a** is accepted. However, the output for the multiple linear regression shows that *reciprocity* is not a significant predictor of total amount funded. This means that **H2b** is rejected. **H3** has been formulated using literature that describes the crowding-out of donators that are monetary compensated for their charitable behaviour in a non-profit setting. The output of model 3 shows that both *reciprocity* and the interaction term *reciprocity*NPO* are insignificant. The findings do not provide enough evidence to accept this hypothesis. Therefore, **H3** is rejected.

H4 is formulated based on findings described in section 2.4. Crowdfunding campaigns allow donators to share campaigns on social media, which facilitates electronic word-of-mouth (e-WOM). Previous research has shown that e-WOM significantly impacts the propensity to pay a higher price for a brand and the intention to purchase. Based on these findings, it is expected that crowdfunding campaigns with high e-WOM are likely to receive more funds than campaigns with low e-WOM. In this thesis, the number of Facebook shares is used as an indicator of e-WOM. In line with **H4**, *shares* is a positive and significant predictor of the total amount funded and is therefore accepted. The final hypotheses focuses on polarity used in campaign descriptions. As explained in section 2.6.1, campaign initiators are able to use positive or negative polarity in their language to describe the crowdfunding campaign. The output of logistic regression shows that sentiment score is an insignificant predictor of campaign success. Thus, **H5a** is rejected. In support of **H5b**, sentiment score is a significant predictor with a positive coefficient. This indicates that campaign descriptions with an overall positive polarity have a positive influence on the total funded amount.

6.1 Discussion and limitations

This thesis focuses on identifying cues used by campaign initiators that might influence the intention to donate to a crowdfunding campaign. Mainly, reciprocity and altruism are discussed as concepts that drive donation behaviour. This thesis adds to the recent stream of crowdfunding literature by applying word embedding algorithms to build lexicons that identify altruistic and reciprocal cues used in campaign descriptions. Also, the effect of differences in polarity used in campaign descriptions on campaign success is analysed for campaigns on GoFundMe. Finally, by using Facebook shares as a proxy of e-WOM, the effect of the amount of campaign shares on campaign success is investigated.

The outcomes of **H1a-2b** show some interesting insights. Although altruistic cues used in campaign descriptions positively influence the total amount funded, the results of logistic regression provided not enough evidence to conclude that altruistic cues significantly impact campaign success. This is reversed for reciprocal cues. Therefore, using reciprocal cues (such as rewarding donators with a gift) in campaign descriptions can help campaign initiators to 'make the campaign happen'. This is especially important for those campaigns that really require the goal amount to be reached in order for the campaign to be realized. Additionally, utilizing altruistic cues in campaign descriptions can significantly increase the total amount of funds received.

The findings for **H4** and **H5b** support some of the findings of previous researchers. First, an increase in the number of shares a campaign receives significantly and positively influences the amount of funds

received. Therefore, facilitating and promoting the opportunity to share the campaign should be stimulated by campaign creators. Finally, using positive polarity in campaign descriptions will positively influence the amount of funds received. Campaign creators should therefore focus on describing the campaign with positive polarity. This could for example be applied for describing the campaign itself, the individual or cause for which the funds are collected and the importance of the campaign.

Some limitations of this thesis have to be mentioned. The first limitation considers the model performance of logistic regression. Classification models tend to favour the most dominant class in predicting campaign success. Because the dependent variable used in this thesis was unevenly distributed (71% of campaigns within dataset is successful versus 29% being unsuccessful), logistic regression faced difficulties in predicting campaigns as successful. Different solutions can be applied to rebalance the data, such as introducing a weighting function that assigns higher weights to unsuccessful campaigns. However, this introduces bias towards the minority class. Another solution to rebalance the data could be to include campaigns that have been closed by the initiator. The reason that these campaigns are closed is most likely because they reached their goal and therefore are no longer in need of collecting funds. GoFundMe does however remove all information of campaigns that are closed by the initiator. For this reason, it is not possible to include these campaigns in the analyses using GoFundMe as a crowdfunding platform. Another issue that was faced is the limited number of campaigns that have been classified as either including altruistic or reciprocal donation cues. Especially campaigns that include reciprocal cues were hard to identify. This is probably due to GoFundMe primarily focusing on non-profit crowdfunding. Other than incorporating weighting functions to assign higher weights to these campaigns, another solution would be to include campaigns from other crowdfunding platforms, such as Kickstarter, who focus more on for-profit crowdfunding (mainly for start-ups). Future research could focus on including and identifying for-profit campaigns. Including these campaigns will probably allow for additional insights as for the differences in using altruistic/reciprocal cues in both for-profit and non-profit crowdfunding. However, including different crowdfunding platforms introduces additional complexity in the analyses, including comparability problems and the need of controlling external influences on campaign success for the platforms. Another limitation regards the usage of word embedding algorithms for building lexicons. Although GloVe in principle is effective and accurate at finding n-grams that are closest related to a specific token, using distance to a token as a metric of similarity introduces a lot of noise when the specific token is not specifically followed by terms that describe it. This problem is described in section 4.5 and requires lexicons to be manually adjusted to remove irrelevant n-grams.

As discussed before, future research could focus on including for-profit campaigns to study additional insights with regards to the differences for altruistic- and reciprocal cues for both forms of crowdfunding. Additionally, as this thesis solely focuses on campaigns established in the United States of America, it could be interesting to study cross-country differences for drivers of campaign success. Finally, the impact of sentiment used in GoFundMe campaigns could be further explored by considering a similar approach as Chen et al. (2016), who focused on identifying different appeal modes and their impact on campaign success.

7. References

Agrawal, A. K., Catalini, C., & Goldfarb, A. (2011). The geography of crowdfunding. National bureau of economic research.

Albert, N., & Merunka, D. (2013). The role of brand love in consumer-brand relationships. *Journal of Consumer Marketing*, *30*(3), 258-266

André, K., Bureau, S., Gautier, A., & Rubel, O. (2017). Beyond the opposition between altruism and self-interest: Reciprocal giving in reward-based crowdfunding. *Journal of Business Ethics*, *146*(2), 313-332.

Andreoni, J. (1989). Giving with impure altruism: Applications to charity and Ricardian equivalence. *Journal of political Economy*, *97*(6), 1447-1458.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The economic journal*, *100*(401), 464-477.

Arrow, K. J. (1972). Gifts and exchanges. *Philosophy & Public Affairs*, 343-362.

Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. (2009). On smoothing and inference for topic models. *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, 27-34.

Axelrod, R. (1987). The evolution of strategies in the iterated prisoner's dilemma. *The dynamics of norms*, 1-16.

Batra, R., Ahuvia, A., & Bagozzi, R. P. (2012). Brand love. *Journal of marketing*, *76*(2), 1-16.

Batson, C. D. (2014). *The altruism question: Toward a social-psychological answer*. Psychology Press.

Belleflamme, P., Lambert, T., & Schwienbacher, A. (2013). Individual crowdfunding practices. *Venture Capital*, *15*(4), 313-333

Belleflamme, P., Lambert, T., & Schwienbacher, A. (2014). Crowdfunding: Tapping the right crowd. *Journal of business venturing*, *29*(5), 585-609.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and economic behavior*, *10*(1), 122-142.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Burkett, E. (2011). A Crowdfunding Exemption? Online Investment Crowdfunding and US Securities Regulation. *Transactions: The Tennessee Journal of Business Law*, *13*(1), 63.

Bowman, D., & Narayandas, D. (2001). Managing customer-initiated contacts with manufacturers: The impact on share of category requirements and word-of-mouth behavior. *Journal of marketing Research*, *38*(3), 281-297.

Brants, T. (2000). TnT: a statistical part-of-speech tagger. *Proceedings of the sixth conference on Applied natural language processing*, 224-231.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth, California: CRC Press.

Buunk, B. P., & Schaufeli, W. B. (1999). Reciprocity in interpersonal relationships: An evolutionary perspective on its importance for health and well-being. *European review of social psychology*, *10*(1), 259-291

Chen, S., Thomas, S., & Kohli, C. (2016). What Really Makes a Promotional Campaign Succeed on a Crowdfunding Platform?: Guilt, Utilitarian Products, Emotional Messaging, And Fewer But Meaningful Rewards Drive Donations. *Journal of Advertising Research*, *56*(1), 81-94

Christidis, K., & Mentzas, G. (2013). A topic-based recommender system for electronic marketplace platforms. *Expert Systems with Applications*, *40*(11), 4370-4379.

Ciresan, D., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2012). Deep neural networks segment neuronal membranes in electron microscopy images. *Advances in neural information processing systems*, 2843-2851.

Clark, S., & Pulman, S. (2007). Combining symbolic and distributional models of meaning. *Proceedings of the AAAI Spring Symposium on Quantum Interaction*, 52–55.

Clavien, C., & Klein, R. A. (2010). Eager for fairness or for revenge? Psychological altruism in economics. *Economics & Philosophy*, *26*(3), 267-290.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1), 15-18.

Cordova, A., Dolci, J., & Gianfrate, G. (2015). The determinants of crowdfunding success: evidence from technology projects. *Procedia-Social and Behavioral Sciences*, *181*, 115-124.

Dabos, G. E., & Rousseau, D. M. (2004). Mutuality and reciprocity in the psychological contracts of employees and employers. *Journal of Applied Psychology*, *89*(1), 52.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013, June). Predicting depression via social media. *Seventh international AAAI conference on weblogs and social media*.

Decety, J. (2015). The neural pathways, development and functions of empathy. *Current Opinion in Behavioral Sciences*, *3*, 1-6.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum

Dellarocas, C., Awad, N., & Zhang, X. (2004). Exploring the value of online reviews to organizations: Implications for revenue forecasting and planning. *ICIS 2004 Proceedings*, 30.

Dellarocas, C., & Narayan, R. (2006). A statistical measure of a population's propensity to engage in post-purchase online word-of-mouth. *Statistical science*, *21*(2), 277-285.

DellaVigna, S., List, J. A., & Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *The quarterly journal of economics*, *127*(1), 1-56.

Dutton, J. E., Dukerich, J. M., & Harquail, C. V. (1994). Organizational images and member identification. *Administrative science quarterly*, 239-263.

Faircloth, J. B. (2005). Factors influencing nonprofit resource provider support decisions: applying the brand equity concept to nonprofits. *Journal of marketing theory and practice*, *13*(3), 1-15.

Fan, Y. W., & Miao, Y. F. (2012). Effect of electronic word-of-mouth on consumer purchase intention: The perspective of gender differences. *International Journal of Electronic Business Management*, *10*(3), 175.

Fehr, E., & Schmidt, K. M. (2006). The economics of fairness, reciprocity and altruism–experimental evidence and new theories. *Handbook of the economics of giving, altruism and reciprocity*, *1*, 615-691.

Filliat, D. (2007). A visual bag of words method for interactive qualitative localization and mapping. *Proceedings 2007 IEEE International Conference on Robotics and Automation*, 3921-3926.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *The annals of statistics*, *28*(2), 337-407.

Frey, B. S., & Meier, S. (2004). Social comparisons and pro-social behavior: Testing "conditional cooperation" in a field experiment. *American Economic Review*, *94*(5), 1717-1722.

Gneezy, U., & List, J. A. (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, *74*(5), 1365-1384.

Godes, D., & Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. *Marketing science*, *23*(4), 545-560.

Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American sociological review*, 161-178.

Griffin, Z. J. (2012). Crowdfunding: fleecing the American masses. *Journal of Law, Technology & the Internet*, *4*(2), 375.

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, *59*, 467-483.

Harms, M. (2007). What drives motivation to participate financially in a crowdfunding community? (Master's thesis). Retrieved from SSRN online database (Accession No. 2269242).

Haski-Leventhal, D. (2009). Altruism and volunteerism: The perceptions of altruism in four disciplines and their impact on the study of volunteerism. *Journal for the Theory of Social Behaviour*, *39*(3), 271-299.

He, Y., & Lai, K. K. (2014). The effect of corporate social responsibility on brand loyalty: the mediating role of brand image. *Total Quality Management & Business Excellence*, *25*(3-4), 249-263.

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent dirichlet allocation. A*dvances in neural information processing systems*, 856-864.

Homburg, C., Ehm, L., & Artz, M. (2015). Measuring and managing consumer sentiment in an online community environment. *Journal of Marketing Research*, *52*(5), 629-641.

Huang, S., Niu, Z., & Shi, C. (2014). Automatic construction of domain-specific sentiment lexicon based on constrained label propagation. *Knowledge-Based Systems*, *56*, 191-200.

Imhof, L. A., Fudenberg, D., & Nowak, M. A. (2007). Tit-for-tat or win-stay, lose-shift?. *Journal of theoretical biology*, *247*(3), 574-580.

Jacobi, C., Van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, *4*(1), 89-106.

Jacobs, B. J., Donkers, B., & Fok, D. (2016). Model-based purchase predictions for large assortments. *Marketing Science*, *35*(3), 389-404.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, *37*(2), 183-233.

Jumpstart Our Business Startups Act of 2017, 15 USC §§301-305.

Keohane, R. O. (1986). Reciprocity in international relations. *International organization*, *40*(1), 1-27.

Kim, C. K., Han, D., & Park, S. B. (2001). The effect of brand personality and brand identification on brand loyalty: Applying the theory of social identification. *Japanese psychological research*, *43*(4), 195-206.

Kim, C., Lee, Y., & Park, B. U. (2001). Cook's distance in local polynomial regression. *Statistics & probability letters*, 54(1), 33-40.

Kim, Y., & Shim, K. (2014). TWILITE: A recommendation system for Twitter using a probabilistic model based on latent Dirichlet allocation. *Information Systems*, *42*, 59-77

Kleemann, F., Voß, G. G., & Rieder, K. (2008). Un(der) paid innovators: The commercial utilization of consumer work through crowdsourcing. *Science, technology & innovation studies*, *4*(1), 5-26.

Lam, P. T., & Law, A. O. (2016). Crowdfunding for renewable and sustainable energy projects: An exploratory case study approach. *Renewable and Sustainable Energy Reviews*, *60*, 11-20.

Leroy, G., Gu, Y., Pettygrove, S., & Kurzius-Spencer, M. (2017). Automated Lexicon and Feature Construction Using Word Embedding and Clustering for Classification of ASD Diagnoses Using EHR. *International Conference on Applications of Natural Language to Information Systems*, 34-37.

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 2177-2185.

Levy, O. & Goldberg, Y. (2014b). Word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Levy, O., & Goldberg, Y. (2014c). Linguistic regularities in sparse and explicit word representations. *Proceedings of the eighteenth conference on computational natural language learning*, 171-180.

Manning, C. D., Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Manning, C., Raghavan, P., & Schütze, H. (2010). Introduction to information retrieval. *Natural Language Engineering*, *16*(1), 100-103.

Marin, L., Ruiz, S., & Rubio, A. (2009). The role of identity salience in the effects of corporate social responsibility on consumer behavior. *Journal of business ethics*, *84*(1), 65-78.

Martin, R., & Randal, J. (2008). How is donation behaviour affected by the donations of others?. *Journal of Economic Behavior & Organization*, *67*(1), 228-238.

Màrquez, L., & Rodríguez, H. (1998). Part-of-speech tagging using decision trees. *European Conference on Machine Learning*, 25-36.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mitra, T., & Gilbert, E. (2014). The language that gets people to give: Phrases that predict success on kickstarter. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 49-61.

Moqri, M., & Bandyopadhyay, S. (2016). Please share! Online word of mouth and charitable crowdfunding. *Workshop on E-Business*, 162-169.

Mohammad, S. M., & Yang, T. W. (2011). Tracking sentiment in mail: How genders differ on emotional axes. *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis*, 70-79.

Mohammad, S. M., & Turney, P. D. (2013). NRC emotion lexicon. *National Research Council, Canada*.

Mollick, E. (2014). The dynamics of crowdfunding: An exploratory study. *Journal of business venturing*, *29*(1), 1-16.

Moro, S., Cortez, P., & Rita, P. (2015). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Expert Systems with Applications*, *42*(3), 1314-1324.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544-551.

Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science*, *43*(1), 88-102.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*(7063), 1291.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *science*, *314*(5805), 1560-1563

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1-135.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532-1543.

Piliavin, J. A., & Charng, H. W. (1990). Altruism: A review of recent theory and research. *Annual review of sociology*, *16*(1), 27-65.

Pitschner, S., & Pitschner-Finn, S. (2014). Non-profit differentials in crowd-based financing: Evidence from 50,000 campaigns. *Economics Letters*, *123*(3), 391-394.

Plisson, J., Lavrac, N., & Mladenic, D. (2004). A rule based approach to word lemmatization. *Proceedings of IS-2004*, 83-86.

Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008). Fast collapsed gibbs sampling for latent dirichlet allocation. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 569-577.

Powers, D. M. (2011). Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technology, 2*(1), 37-63.

Purver, M., Griffiths, T. L., Körding, K. P., & Tenenbaum, J. B. (2006). Unsupervised topic modelling for multi-party spoken discourse. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 17-24.

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 46-50.

Reisenbichler, M., & Reutterer, T. (2019). Topic modeling in marketing: recent advances and research opportunities. *Journal of Business Economics*, *89*(3), 327-356.

Reza Jalilvand, M., & Samiei, N. (2012). The effect of electronic word of mouth on brand image and purchase intention: An empirical study in the automobile industry in Iran. *Marketing Intelligence & Planning*, *30*(4), 460-476.

Rose-Ackerman, S. (1996). Altruism, nonprofits, and economic theory. *Journal of economic literature*, *34*(2), 701-728.

Rose-Ackerman, S. (1997). Altruism, ideological entrepreneurs and the non-profit firm. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, *8*(2), 120-134.

Ryan, J., & Van Wingerden, R. P. (2011). Fighting for Funds: An Exploratory Study into the Field of Crowdfunding. *Extraction, 14*(151), 1-82.

Sargeant, A. (1999). Charitable giving: Towards a model of donor behaviour. *Journal of marketing management*, *15*(4), 215-238.

Schwienbacher, A., & Larralde, B. (2010). *Crowdfunding of small entrepreneurial ventures*. Handbook of entrepreneurial finance, Oxford University Press, Forthcoming.

Sen, S., & Bhattacharya, C. B. (2001). Does doing good always lead to doing better? Consumer reactions to corporate social responsibility. *Journal of marketing Research*, *38*(2), 225-243.

Settoon, R. P., Bennett, N., & Liden, R. C. (1996). Social exchange in organizations: Perceived organizational support, leader–member exchange, and employee reciprocity. *Journal of applied psychology*, *81*(3), 219.

Shehu, E., Becker, J. U., Langmaack, A. C., & Clement, M. (2016). The brand personality of nonprofit organizations and the influence of monetary incentives. *Journal of Business Ethics*, *138*(3), 589-600.

Silverman, G. (2001). *The Secrets of Word-of-Mouth Marketing: How to Trigger Exponential Sales Through Runaway Word-of-Mouth*. New York: American Marketing Association.

Sober, E., & Wilson, D. S. (1999). *Unto others: The evolution and psychology of unselfish behavior*. Harvard University Press.

Stanko, M. A., & Henard, D. H. (2017). Toward a better understanding of crowdfunding, openness and the consequences for innovation. *Research Policy*, *46*(4), 784-798.

Stewart-Williams, S. (2007). Altruism among kin vs. nonkin: effects of cost of help and reciprocal exchange. *Evolution and human behavior*, *28*(3), 193-198.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, *427*(7), 424-440.

Sugden, R. (1982). On the economics of philanthropy. *The Economic Journal*, *92*(366), 341-350.

Sugden, R. (1984). Reciprocity: the supply of public goods through voluntary contributions. *Economic Journal*, 94(376), 772-787.

Sung, Y., & Kim, J. (2010). Effects of brand personality on brand trust and brand affect. *Psychology & Marketing*, *27*(7), 639-661.

Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*, *51*(4), 463-479.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly review of biology*, *46*(1), 35-57.

Variable importance (2019). Retrieved from https://topepo.github.io/caret/variable-importance.html

Venable, B. T., Rose, G. M., Bush, V. D., & Gilbert, F. W. (2005). The role of brand personality in charitable giving: An assessment and validation. *Journal of the academy of marketing science*, *33*(3), 295-312.

De Waal, F. B. (2008). Putting the altruism back into altruism: the evolution of empathy. *Annual Review of Psychology*, *59*, 279-300.

Wang, W., Zhu, K., Wang, H., & Wu, Y. C. J. (2017). The impact of sentiment orientations on successful crowdfunding campaigns through text analytics. *IET Software*, 11(5), 229-238.

Warneken, F., & Tomasello, M. (2009). The roots of human altruism. *British Journal of Psychology*, *100*(3), 455-471.

Wei, X., & Croft, W. B. (2006). LDA-based document models for ad-hoc retrieval. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 178-185.

Wendlandt, L., Kummerfeld, J. K., & Mihalcea, R. (2018). Factors influencing the surprising instability of word embeddings. *arXiv preprint arXiv:1804.09692*.

White, K., & Peloza, J. (2009). Self-benefit versus other-benefit marketing appeals: Their effectiveness in generating charitable support. *Journal of Marketing*, *73*(4), 109-124.

Wilson, D. S. (1992). On the relationship between evolutionary and psychological definitions of altruism and selfishness. *Biology and Philosophy*, *7*(1), 61-68.

Wu, L., Hoi, S. C., & Yu, N. (2010). Semantics-preserving bag-of-words models and applications. *IEEE Transactions on Image Processing*, *19*(7), 1908-1920.

Yuan, H., Lau, R. Y., & Xu, W. (2016). The determinants of crowdfunding success: A semantic text analytics approach. *Decision Support Systems*, *91*, 67-76.

Zvilichovsky, D., Inbar, Y., & Barzilay, O. (2015). Playing both sides of the market: Success and reciprocity on crowdfunding platforms. Retrieved from SSRN online database (Accession No. 2304101).

Zvilichovsky, D., Danziger, S., & Steinhart, Y. (2018). Making-the-Product-Happen: A driver of crowdfunding participation. *Journal of Interactive Marketing*, *41*, 81-93.

Appendix A – Latent Dirichlet Allocation

Figure 8.1: Perplexity plot for number of topics (k) between 5 – 60.

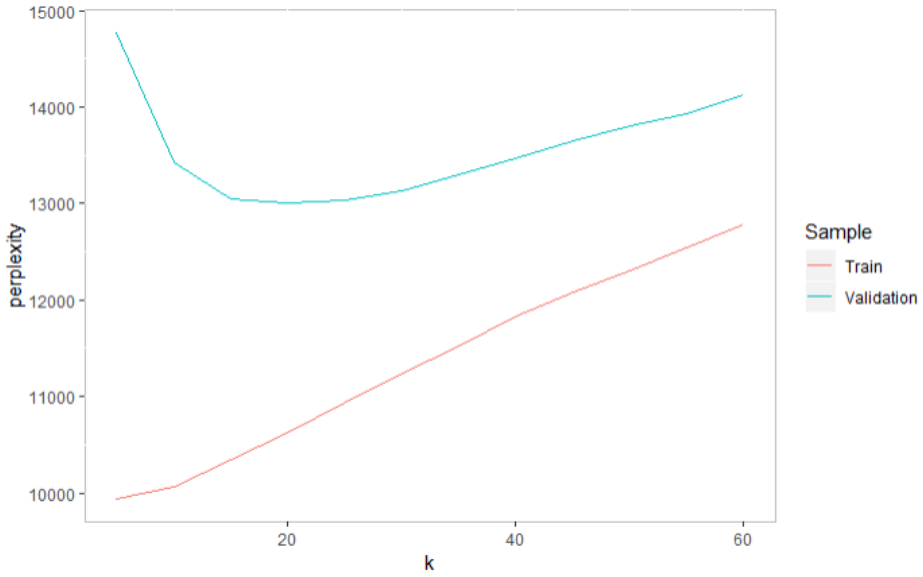


Figure 8.2: Perplexity plot for alpha between 0.01 – 0.2.

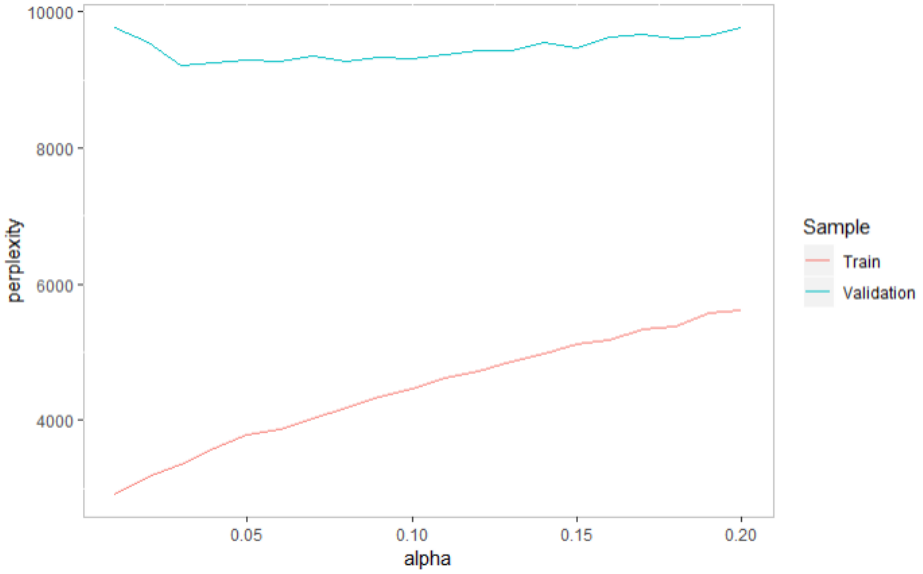
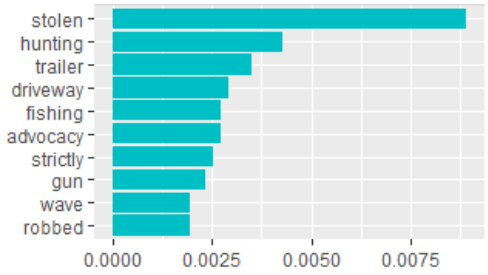
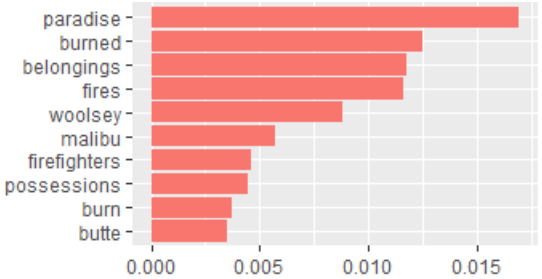


Figure 8.3: Top-10 terms per topic based on per-topic word distribution

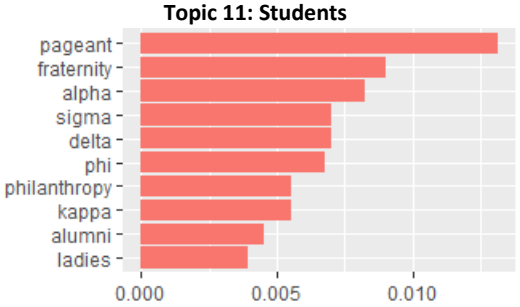
Topic 5: Crime-related



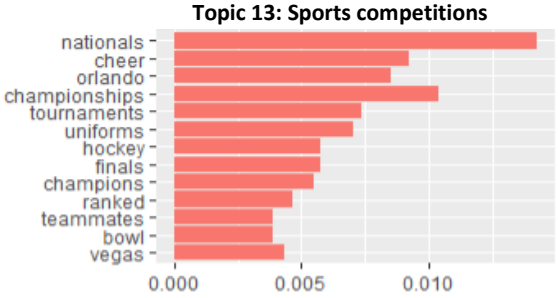
Topic 8: Natural disasters



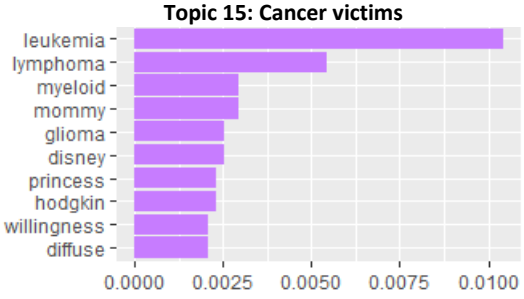
Topic 11: Students



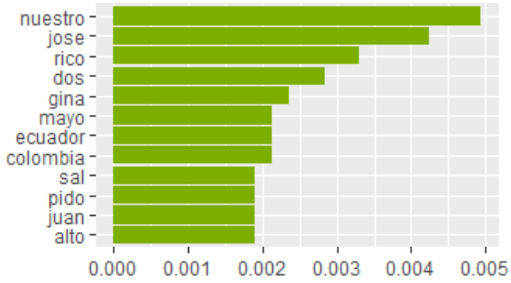
Topic 13: Sports competitions



Topic 15: Cancer victims



Topic 20: Hispanic language



Appendix B – Correlation plots

Figure 8.4: Correlation plot of all predictor variables considered for logistic regression.

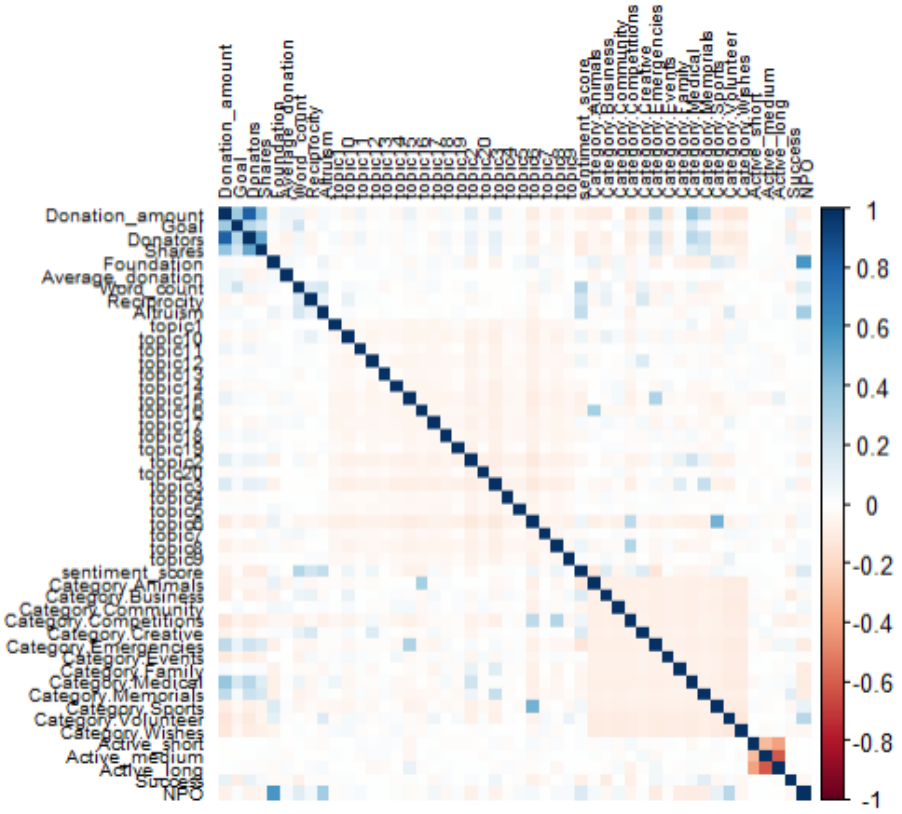


Figure 8.5: Correlation plot of "active".

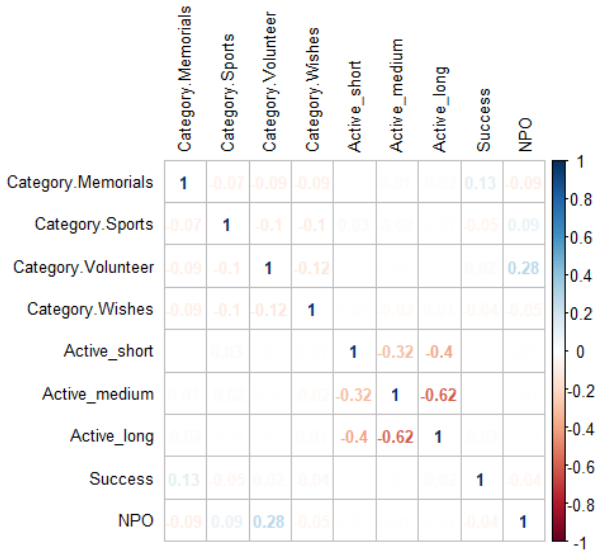


Figure 8.6: Correlation plot of numerical variables.